Measuring cognitive state from physiological signals in user interface research

Kseniia Belorustceva

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 6.1.2018

Supervisor

Prof. Antti Oulasvirta

Advisor

D.Sc. Ilkka Nissilä

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Kseniia Belorustceva

Title Measuring cognitive state from physiological signals in user interface research

Department Department of Neuroscience and Biomedical Engineering

Major Human Neuroscience and Technology **Code of major**

Supervisor Prof. Antti Oulasvirta

Advisor D.Sc. Ilkka Nissilä

Date 6.1.2018 **Number of pages** 4+46 **Language** English

Abstract

The purpose of this Thesis is to investigate how modern technology can be used for evaluating human cognitive state in the context of human-computer interaction, namely user interface (UI) research.

In this work two types of physiological data were collected to measure cognitive load during a task which requires some degree of human-computer interaction. A near-infrared spectroscopy device and eye tacker were used to evaluate cognitive load level during the task and provide an insight into how these data might be used in an adaptive real-time system. A mental calculation task was used as the cognitively demanding learning task, which challenges working memory. Additional difficulty was added using the task presentation: mathematical expressions were either static or moving from the top to the bottom of the screen.

Results indicate that tasks of different mental complexity elicit different cognitive responses. With careful interpretation this information can be used in designing environments, suitable for the user.

This work have shown that in designing the systems which use physiological measurements, it is crucial to know the possible sources of the noise. For example, in pupillary measurements it is important to control for luminance and physiological changes which affect pupil size along with cognitive load, or to develop methods which discriminate between task-evoked pupil response from other responses.

For any real-time system it is necessary to develop the fast and efficient algorithms which produce reliable results with minimal training of the models.

Keywords Cognitive load, hemodynamic brain imaging, task-evoked pupillary response, user interfaces

Preface

This thesis was created between July and December 2017 in the Department of Neuroscience and Biomedical Engineering of Aalto University.

First I would like to thank my advisor Ilkka Nissilä and members of his research group for supporting my ideas with the state-of-art neuroimaging technology. It was a pleasure to be the part of the research process.

I am thankful to my supervisor Antti Oulasvirta for giving me freedom to explore the topic and a guidance when I needed it.

I am happy to say that moving to study in Aalto University was made possible only with the support of my parents, to whom I am infinitely grateful.

Finally, I want to thank my husband Matti Palin for continuous support and motivation throughout my studies.

Espoo, 6 January 2018

Kseniia Belorustceva

Contents

Symbols and abbreviations

Abbreviations

ACT	Adaptive control of thought
AG	Angular gyrus
BCI	Brain-computer interface
BP	Blood pressure
CL	Cognitive load
CLT	Cognitive load theory
CT	Computed tomography
EEG	Electroencephalography
fMRI	Functional magnetic resonance imaging
GSR	Galvanic skin response
HCI	Human-computer interaction
IFG	Inferior frontal gyri
IPS	Intra-parietal sulcus
IS	Information systems
LTM	Long-term memory
MEG	Magnetoencephalography
MFG	Middle frontal gyri
MIRS	Near-infrared spectroscopy
PET	Positron emission tomography
PPS	Posterior parietal cortex
ROI	Region of interest
SFG	Superior frontal gyri
SPL	Superior parietal lobe
UI	User interface
WM	Working memory

1 Introduction

The area of user interface engineering aims to make the user's interaction with the UI as easy and efficient as possible in terms of accomplishing user goals. A good interface design facilitates finishing the task without drawing unnecessary attention to itself. Interface design is involved in a wide range of projects in almost every area of science.

User interface design requires understanding of user psychology, physiology and intentions. An adaptive user interface is an interface that changes according to user needs or context, thus it has an advantage in terms of user satisfaction and performance. One of the most complicated issues in designing an adaptive interface is predicting what change will give the user better understanding and improved efficiency during the task at hand without interfering with user experience. Researchers in the field of human-computer interaction (HCI) observe the ways in which humans interact with computers and develop technologies which allow this interaction to happen in novel ways. Many scientific fields intersect in the HCI research: behavioral science, neuroscience, computer science, software engineering and design, etc.

The potential of neuroscience for HCI research have been increasing due to increased availability of neuroscience methods and theories. Advances in brain imaging technologies and cognitive neuroscience provide the ability to interface directly with the human brain. Such connection can be implemented through sensors which can monitor some of physiological processes that occur in the brain and correspond with certain forms of thought. Researchers might utilize brain imaging technologies to build Brain-Computer Interface (BCI), a system which does not depend on brain's normal output pathways of peripheral nerves and muscles [1]. With BCI tools, researchers in HCI can explore conceptually new forms of human-computer interaction: for example interfaces which do not require explicit input in order to direct the computer to do something, but, rather, can infer information about user state from user physiology, behavior or the environment. With this information the systems can dynamically adapt themselves in order to support the user in the task at hand.

Researchers in the field of Information Systems (IS) have begun to investigate the value of neuroscience for IS research, creating a new field named NeuroIS [2]: "NeuroIS is a subfield in the IS literature that relies on neuroscience and neurophysiological theories and tools to better understand the development, use, and impact of information technologies (IT)." NeuroIS seeks to contribute to the development of new theories that enable accurate predictions of IT-related behaviors, and to the design of IT artifacts that positively affect economic and noneconomic variables (e.g., productivity, satisfaction, adoption, well being). Clearly, neuroscience theories and tools can significantly contribute to HCI, BCI, design and decision-making.

The necessary steps towards designing and developing adaptive user interfaces are understanding the user's cognitive processes involved in the interaction with the interface and knowing means by which the interface can influence user's performance. In addition, given an increasing importance of neuroscience tools among HCI researchers, it is crucial to investigate whether non-invasive techniques allow to study cognitive processes with precision, comparable to more traditional neuroimaging techniques.

Motivations of the work include the following points.

- Understanding of the human cognition is crucial for developing high quality user interfaces, especially adaptive ones.
- Although human cognition can be modeled to some extent, real-time physiological data might provide additional knowledge about how human cognitive state changes in response to the environment, and importantly, how to change the environment in order to elicit certain cognitive response.
- Among wide range of neuroimaging and other physiological devices available nowadays only few are suitable for non-medical research settings, such as user interface research.
- Cognitive state of a user in UI research is a product of complex influences. It is important to understand whether subtle changes in cognitive state induced by certain aspects of user interface can be measured with available devices.

2 Thesis goals and tasks

The goal of this Thesis is to show how methods in human neuroscience and technology can be used for user interface research, namely evaluating human cognitive state in the context of human-computer interaction. The task is to investigate how an adaptive user interface can benefit from the information about user's cognitive state, and whether this information can be inferred from physiological measurements of hemodynamic response and pupillary response. In this work the focus is in one aspect of the cognitive state of the user, namely the cognitive load (CL). The user interface is a computer screen with a cognitively demanding task presented on it. Definition of the cognitive load and the choice of the cognitive task are described in detail in the section Background.

In order to build a simple brain-computer interface one has to deal with the following tasks:

- Express the cognitive state as a physiological signal. This requires definition of a representation of the cognitive state of the user and verifying that such representation is supported by existing models or data.

  In this work, cognitive state shall be represented through cognitive load. Physiological data and behavioral data is used to measure CL and to estimate individual differences in the influence of cognitively demanding task to the CL.

- Manipulate cognitive state. When developed models or acquired data can reliably explain changes in cognitive state in response to the cognitively demanding task, one can manipulate the task in order to elicit different cognitive responses.

  In our context this means manipulating one aspect of the task presentation, namely, the speed of vertical movement of the task.

- Use knowledge of cognitive state for task at hand. In reality, cognitive state of the user has a direct influence on user's performance on a given task. The complexity of the environment (e.g. distractions of divided attention) brings additional difficulty, but the reward for achieving the goal might increase if the goal was achieved under difficult conditions.

  In this work the task of the user is to accurately perform mental calculations, and additional difficulty is introduced by the task presentation.

- Real-time BCI. The usefulness of the real-time BCI is hard to overestimate. In order to achieve good real-time performance one has to constantly improve models and algorithms of the system and overcome a number of limitations. Some of such limitations have been investigated in the work by [3].

This thesis is organized to sections as follows. Section Background gives an overview of the related research in the field of human-computer interaction along with the description of the physiological tools and theories used in this work. Human cognitive architecture, physiological measurements and tools, as well as mental arithmetics are discussed. Section Experiment overview introduces the user pilot study. Section Materials and methods describes the scientific experiment conducted in order to evaluate the designed system, as well as data collection and analysis. Section Discussion provides the detailed assessment of the results. Section Future work addresses current limitations and ways to overcome them.

3 Background

One of the biggest challenges facing brain-computer interface researchers today is the basic mechanics of the interface itself. Nowadays researchers mainly use non-invasive brain imaging techniques such as functional magnetic resonance imaging (fMRI), computed tomography (CT), positron emission tomography (PET), electroencephalography (EEG), magnetoencephalography (MEG), near infrared spectroscopy (NIRS). Other physiological measures include blood pressure, body temperature, galvanic skin response (GSR), and size of the pupil.

EEG and MEG measure changes in magnetic fields at the scalp caused by changing electrical currents in brain neurons. The main strength of these techniques is their millisecond-level time precision. Brain imaging techniques based on the brain's consumption of glucose (via PET scanning) or oxygen (via functional MRI) provide a more delayed response to cognition but enable 3D localization of brain activity with millimeter-level spatial resolution, which has led to their widespread use in functional neurology and neuroanatomy. For many years fMRI along with EEG are the primary neuroimaging modalities for cognitive neuroscience. Both methods, although providing good spacial and temporal resolution, are very sensitive to artifacts such as movement or blinking, and require complex equipment.

In neuroscience research and in medical settings any concrete method or combination of methods can be chosen freely according to the research objective or the environment. In contrast, equipment for HCI experiment should be the least complex, allowing easy setup and some interaction of the subject with the external world. The choice of the physiological measurement tools is discussed in the subsection Physiological measurements and tools.

It is important to note that physiological measurements, such as cortical oxygen consumption, pupillary response, electrical activity of the brain, etc., provide only indirect measurement of processes in the brain. Researchers are using knowledge of human biology, anatomy and models of human cognition in order to explain these processes in terms of cognition. Therefore, it is important to describe the human cognitive architecture as it is known nowadays. High-level overview of human cognition is given in subsection Human cognitive architecture.

In BCI/HCI user experiments the tasks must be carefully thought out: the task execution by the user must initiate certain cognitive processes, which, in turn, should cause an activation in certain brain regions. Finally, it should be possible to infer from the physiological measurements, that activation was indeed caused by the task. The choice of such task discussed in subsection Mental arithmetics.

Past research of Fishburn [4], Hossain [5], Klingner [6] has shown that cognitive load levels correlate with the cognitive task difficulty. Keshmiri [7] introduces non-parametric approach to estimate CL using NIRS time series. In their study Fishburn [4] and Keshmiri [7] used n-back task and fNIRS for data acquisition; Hossain [5] inferred cognitive load from task-evoked pupillary response on a visually presented mental calculation task; Klingner [6] successfully replicated several pupillometry studies and addressed several limitations of cognitive pupillometry, enabling the use of trial-averaged pupillometry to measure cognitive load during simple visual tasks.

Verner [8] showed, using NIRS, that a cortical oxygen consumption is a function of task difficulty in mental arithmetic task.

3.1 Human cognitive architecture

In humans, the parts of the brain responsible for encoding sensory information and commanding movements (i.e. primary sensory and motor cortices) account for only a fraction of the cerebral cortex. The remaining cortex is concerned with focusing on complex stimuli, identifying the relevant features of such stimuli, recognizing the related objects, and planning appropriate responses, as well as storing aspects of this information [9]. Collectively, these integrative abilities are referred to as cognition, and, evidently, cortices in parietal, temporal and frontal lobes are parts of the brain which make cognition possible. These areas of the cerebral cortex are collectively referred to as the association areas. One of the most intriguing of the brain's complex functions is the ability to store information provided by experience and retrieve it when necessary. Learning is the process by which new information is acquired by the nervous system and is observable through changes in behavior: the acquisition of new knowledge or skills. Memory refers to encoding, storage and retrieval of learned information.

Human brain and mind have not been explored completely, but various results of neuroscience and psychology research allow to theorize and model human cognition to some extent. Cognition has been modeled and studied for the last forty years and the number of existing cognitive architectures is several hundreds. A cognitive architecture specifies the underlying infrastructure for an intelligent system. The broad overview of the cognitive architectures and their practical applications present in a work by Kotseruba et al. [10]. They identified ten major categories of applications, namely human performance modeling, games and puzzles, robotics, psychological experiments, natural language processing, human-robot and human-computer interaction, computer vision, categorization and clustering, virtual agents and miscellaneous. Most of the cognitive architectures include the core cognitive abilities, such as perception, attention mechanisms, action selection, memory, learning and reasoning.

As mentioned earlier, in this work physiological data is used to access task-evoked cognitive load, which, in turn, is used to change the task presentation, i.e. adapt the interface to help the user achieve better performance. In general, such systems are the primary focus of instructional design in the field of educational psychology. With learning theory as a foundation, instruction can be structured around making learning most effective. There exists a number of models describing learning, commonly categorized as behaviorist models, cognitivist models and constructivist models. Two learning theories, which had the most influence on this work, are discussed in this section: Cognitive load theory by John Sweller [11] and ACT-R theory by John Anderson [12].

3.1.1 Human memory

Both of the above mentioned theories, as well as most of other cognitive architectures, rely heavily on the representation of the human memory. The study of the human memory begins from attempts of ancient philosophers to understand the human mind. In antiquity, it was generally assumed that there are two sorts of memory: the inborn "natural memory" and the acquired "artificial memory". In 18th century it was hypothesized that memories were encoded through "hidden motions in the nervous system". In the mid-1880s the German philosopher Herman Ebbinghaus [13] developed the first scientific approach to studying memory, and his findings include classification of the three distinct types of memory: sensory, short-term and long-term, which remain relevant to this day. In 1949 Canadian Donald Hebb intuited that “neurons that fire together, wire together”, implying that the encoding of memories occurs when connections between neurons established through repeated use [14]. This claim, nowadays referred to as Hebb’s Rule, was supported by the discovery of the mechanics of memory consolidation, long-term potentiation and neural plasticity in the 1970s, and remains the reigning theory today. In 1968, Richard Atkinson and Richard Shiffrin first described their model of memory, consisting of a sensory memory, a short-term memory and a long-term memory, which became the most popular model for studying memory for many years [15]. In 1974, Alan Baddeley and Graham Hitch proposed their model of working memory, which consists of the central executive, visuo-spatial sketchpad and phonological loop as a method of encoding [16]. During the 1980s and 1990s, several formal models of memory were developed that can be run as computer simulations, including various versions of the Adaptive Control of Thought (ACT) model developed by John Anderson in 1983 [12].

Nowadays it is generally accepted by researchers in cognitive neuroscience that humans have at least two qualitatively different systems of information storage (generally referred to as declarative and nondeclarative memory) and at least three temporally different systems (immediate memory, working memory, long-term memory) [9].

Declarative memory is the storage and retrieval of material that is available to the consciousness and can be expressed by language. Three famous cases of amnesia, H.M., N.A., and R.B., provided evidence for the importance of midline diencephalic and medial temporal lobe structures (particularly, the hippocampus) in establishing new declarative memories. A large amount of evidence implies that the cerebral cortex is the major long-term repository for many aspects of declarative memory.

Nondeclarative (procedural) memory involves skills and associations that are acquired and retrieved at an unconscious level. Nondeclarative memory is also frequently called implicit memory, because it results from direct experience, and declarative memory is often called explicit memory, because it results from more conscious effort. Nondeclarative memory apparently involves the basal ganglia, prefrontal cortex, amygdala, sensory association cortex, and cerebellum.

Immediate (sensory) memory is the routine ability to hold ongoing experiences in mind for fractions of a second. The capacity of sensory memory is very large and each sensory modality appears to have its own memory register.

Short-term (working) memory is the ability to hold and manipulate information in mind for seconds to minutes and has very limited capacity (known as memory span, generally 5 to 9 items). The central executive part of the prefrontal cortex appears to play a fundamental role in working memory. The central executive controls two neural loops, one for visual data (which activates areas near the visual cortex of the brain and acts as a visual scratch pad), and one for language (the phonological loop, which uses Broca's area as a kind of inner voice that repeats word sounds to keep them in mind). These two scratch pads temporarily hold data until it is erased by the next job.

The long-term memory is the ability to retain information permanently for days, weeks or a lifetime. According to the standard model of memory consolidation [18], information comes through neocortex areas associated with sensory systems and is then sent to the medial temporal lobe for processing (especially the hippocampal system). Changes in synapses create a memory trace via a process called synaptic consolidation. After synaptic consolidation, or perhaps overlapping with it in time, systems consolidation occurs in which engrams (the physical embodiments of the long-term memory in neuronal machinery) are moved gradually over time into distributed areas of the neocortex. Permanent engrams are stored in a variety of neocortical areas. Before systems consolidation, memory retrieval requires the hippocampus, but after systems consolidation is complete, the hippocampus is no longer needed.

3.1.2 Cognitive load theory (CLT)

Nearly thirty years of research supports the connection between deeper levels of semantic processing, or cognitive engagement, and increased learning [19]. Cognitive load (CL) refers to the amount of mental effort being used in the working memory.

Cognitive load theory (CLT) suggests that learning happens best under conditions that are aligned with human cognitive architecture. CLT was developed out of the study of problem solving by John Sweller in the late 1980s [11]. Recognizing George Miller's information processing research [20] showing that short term memory is limited in the number of elements it can contain simultaneously, Sweller argued that instructional design can be used to reduce cognitive load in learners. According to CLT, individual learning depends on the limited processing capacity of the learner's cognitive architecture and CL imposed by the task.

Within CTL framework the cognitive architecture consists of an effectively unlimited long-term memory (LTM), which interacts with a working memory (WM) that is very limited in both capacity and duration. LTM consists of cognitive schemas that store and organize knowledge by incorporating multiple elements of information into a single element, or chunk, with a specific function. CLT effectively deals with the limitations that are induced by WM by creating instructions that lower intrinsic (content-based), extraneous (presentation-based) and germane (information consolidation-based) cognitive load on WM [21]. The load is considered to be intrinsic if it is imposed by the number of information elements in a task and the interactivity between those elements. When the load is imposed by the manner in which information is presented to learners and by learning activities required of them, it is called either extraneous or germane. Extraneous load is imposed by information and activities that do not directly contribute to learning (e.g. searching for the information), while germane load is caused by information and activities that foster the learning process.

Intrinsic, extraneous and germane cognitive load are considered to be additive, so that total load can not exceed the total WM capacity in order for learning to occur [21]. Intrinsic load provides a base load that is irreducible by instructional design. It can only be lowered by constructing additional schemata and automating previously acquired schemata. Any available WM capacity remaining after resources have been allocated to deal with intrinsic load is allocated to extraneous or germane load. A reduction in extraneous load by using a more effective instructional design can free capacity for an increase in germane load. This provides the basis to form the new cognitive schema, which results in reduction in intrinsic load and freeing WM capacity for using the newly learned material to acquire more advanced schemata.

Paas and Van Merriënboer [22] developed a way to measure perceived mental effort which is indicative of cognitive load. According to them, cognitive load is a multidimensional construct representing the load that is imposed on a learner's cognitive system when executing a certain task. It can be conceptualized in two dimensions: a task-based dimension (mental load) and a learner-based dimension (mental effort). Both dimensions affect the learner's performance. Thus, cognitive load needs to be accessed by measuring mental load, mental effort and performance.

Cognitive load can be accessed and monitored using subjective (self-reports) and objective (eye tracking, physiological, task- and performance-based) methods. Subjective measures assume that people are able and willing to monitor and report the amount of mental effort that they invested during task performance. It gives no direct insight into which type of load was demanding the learner's cognitive capacities. Physiological techniques assume that changes in cognitive states evoke physiological changes. These techniques visualize the detailed trend and pattern of load, but can be uncomfortable for the participants and often require complex instrumentation. Task- and performance-based methods can be either the performance of the learned task itself or performance on the secondary task which goes simultaneously with the first task. These measures provide a direct estimate of mental effort while performing a task, but they can affect the primary task itself.

Sweller's theories are best applied in the area of instructional design of cognitively complex or technically challenging material. His focus is on the reasons why people have difficulty learning material of this nature. Cognitive load theory has many implications in the design of learning materials which must, if they are to be effective, keep cognitive load of learners at a minimum during the learning process.

In this work CLT is used to represent the cognitive load elicited by the task as a result of two processes: intrinsic, i.e. devoted to solving the cognitive task itself, and extraneous, i.e. devoted to cope with difficulty added by the task presentation.

3.1.3 Adaptive control of thoughts – rational theory (ACT-R)

ACT (Adaptive Control of Thought) is a cognitive architecture based on the assumption of a unified theory of mind. The theory was mainly developed by John Anderson

[12], beginning from 1983. The goal of this cognitive theory is to explain how human cognition works and what the structures and processes of human memory, thinking, problem solving, and language are. The core of ACT is a production system with a pattern matcher that works on memory and perceptual-motor modules via buffers. The current version, adaptive control of thought - rational (ACT-R) is based on the principle of rationality of the human mind. ACT-R is a cognitive architecture and a theory for simulating and understanding human cognition. Simulations with ACT-R allow for predicting typical measures in psychological experiments like latency (time to perform a task), accuracy (correct vs. false responses), and neurological data (e.g., FMRI-data).

ACT-R distinguishes among three types of memory structures: declarative, procedural and working memory. According to ACT-R, all knowledge begins as declarative information; procedural knowledge is learned by making inferences from already existing factual knowledge. ACT-R supports three fundamental types of learning: generalization, in which productions become broader in their range of application, discrimination, in which productions become narrow in their range of application, and strengthening, in which some productions are applied more often. New productions are formed by conjunction or disjunction of existing productions.

ACT-R can explain a wide variety of memory effects as well as account for higher order skills such as geometry proofs, programming and language learning. As the research continues, ACT-R evolves ever closer into a system which can perform the full range of human cognitive tasks: capturing in great detail the way we perceive, think about, and act on the world.

3.2 Physiological measurements and tools

Experimental settings for user studies, which involve physiological measurements to explore high cognitive functions, must be carried out in realistic environments. In HCI and UI research such environments include interaction with a computer, which makes some neuroimaging methods impractical to use. Research [4], [7], [5], [8] has shown the usefulness of NIRS and eye tracking for studying processes which influence cognitive load. Experimental settings with NIRS and eye tracker are relatively natural for studies involving interaction with the computer screen, especially compared with fMRI and MEG. Therefore, physiological data in this work were acquired in the form of hemodynamic brain response (using NIRS) and pupillary response (using an eye tracker).

3.2.1 Task-evoked pupillary response

Pupil dilation is primarily the result of the integrated activity of two groups of muscles located in the iris. The circular muscles encircle the pupil; when activated, this set serves to constrict the diameter of the pupil and make it smaller. The radial muscles lie immediately outside the circular muscles and extends radially from the pupil out through the iris. When activated, the radial muscles pull the pupil diameter outward and cause it to become larger. Circular and radial muscles typically work together through reciprocal innervation, a physiological process involving both agonistic and antagonistic responses [23].

In the presence of steady light, an individual's pupil responds with a continual but irregular oscillation, known as the light reflex. During the light reflex, the circular muscles act as the agonist and are stimulated to contract, while the radial muscles act as the antagonist and are inhibited from dilating the pupil. The reflex is fleeting, and the result is a visible pulsing of the pupil. When individual experiences a psychosensory stimulus, e.g. a task requiring significant cognitive processing, the pupil may make a response that is quite different from the light reflex as the process of reciprocal innervation is reversed: the radial muscles are activated, causing the pupil to dilate, and the circular muscles are inhibited, also causing the pupil to dilate. The result is the brief dilation that is greater than either muscle group alone could effect. This phenomenon is called the dilation reflex. The fundamental problem in studying the relationship between cognitive activity and pupillary response lies in how to separate the dilation reflex from the light reflex. Both phenomena can occur at the same time, and it should be considered during the experimental research.

Task-evoked pupillary response is caused by a cognitive load imposed on a human and as a result of the decrease in parasympathetic activity in the peripheral nervous system. The tasks place some demand on the working memory. When the demand increases, task-evoked pupillary response results in a linear increase in pupil dilation. This can reflect differences in processing load within a task, between different tasks and between individuals. Task-evoked pupillary response is used as an indicator of cognitive load levels in psychophysiology research and information systems research. Given the cognitive task-evoked pupillary responses, effects of cognitive load can be estimated from the frequency analysis of pupillary time series. However, the temporal changes of pupillary response are nonstationary, nonlinear, and commonly occur with temporal discontinuities. One approach to achieve the desired robustness was suggested by [5]. The Hilbert transform method for the study of cognitive overload and cognitive dissonance reveals new insight into the association between task variation and pupillary dynamics. Some improvements in methods for measuring cognitive load using pupillary dilations are presented in work [6].

3.2.2 Near-infrared spectroscopy

Near-infrared spectroscopy (NIRS) is a neuroimaging technique for recording cortical hemodynamic activity. The oxygen level is associated with increased brain activity, therefore measured blood oxygenation levels in the brain provide evidence of brain activity associated with the task performance. In non-medical settings, such as user interface research, near-infrared spectroscopy is one of the most promising methods, because it does not require a complex setup, it is not as sensitive to artifacts as other methods, and it is non-invasive. The method projects near-infrared light through the scalp and records optical density fluctuations resulting from metabolic changes within the brain. Similar to fMRI, cerebral blood flow is used as a proxy for neuronal activity. Both spatial resolution and penetration depth of fNIRS are dependent on the distances between light sources and detectors. The technique is particularly resilient to contamination from head motion since the optodes are affixed to the head and thus move with the subject [4].

Until recently, it was not clear whether fNIRS is sensitive to load-dependent working memory changes in activation and functional connectivity in prefrontal-parietal regions and whether fNIRS is sensitive to functional connectivity differences between a working memory task and a task-free resting state. The study [4] showed that fNIRS has the required sensitivity to detect activation and functional connectivity that increase linearly with increasing working memory load, and demonstrated that fNIRS can reliably detect differences related to the cognitive state (e.g. rest versus task). Another study [25] showed several workload comparisons with promising levels of classification accuracy.

3.3 Mental arithmetic

Basic mental calculations are essential and support us in many everyday life decisions. Mental arithmetic is a central aspect of mathematical achievement in a primary school, and it is a learned skill. Therefore, mental arithmetic provides an excellent framework for the investigation of fundamental cognitive processes beyond abstract problem solving, such as retrieving information, execution of control processes, and updating the information.

Thinking about the complexity of a mental arithmetic task, one could heuristically assume that task complexity depends on the type of operator and the magnitude of operands. Following that assumption, some studies have interpreted differences in brain activation between arithmetic operation types as evidence in favor of distinct cortical representations, processes or neural systems. Different involvement of visual-spacial, sensory-motor, and verbal processes has been suggested for addition, subtraction and multiplication, e.g. in [24], [17], [26]. Investigation of procedures and strategies underlying problem solving strategies received less attention.

Research in the fields of mathematics education and cognitive psychology revealed that different strategies might be used in solving elementary arithmetic problems (e.g. [8], [29], [28]). These problems are solved either by directly retrieving the correct answer from long-term memory (retrieval strategy) or by using more complex procedural strategy, which involves mix of retrieval, decomposition, counting, and updating processes. Neurophysiological studies have begun to investigate brain correlates of arithmetic strategy use. In the majority of these studies strategy use was examined by systematically varying the size of the operands with the assumption that different problem size would trigger different types of strategies [28].

Study by Rosenberg-Lee [27] examined functional overlap and dissociations in intraparietal sulcus (IPS), superior parietal lobule (SPL) and angular gyrus (AG), across four operations: addition, subtraction, multiplication, and division. Their findings demonstrated that individual IPS, SPL and AG subdivisions are differentially modulated by the four arithmetic operations, pointing to significant functional heterogeneity and individual differences in activation and deactivation within the posterior parietal cortex (PPC). Rosenberg-Lee attributed these affects to retrieval, calculation and inversion, the three key cognitive processes that are differentially engaged by arithmetic operations. Their findings indicate that, compared to a number identification control task, all operations except addition, showed a consistent activation if the left posterior IPS and deactivation in the right posterior AG. No PPC regions showed significant activation during addition, however, significant deactivations were detected in the right AG and adjoining supramarginal gyrus.

A study by Tschentscher [29] contrasted procedural strategies with retrieval strategies with varied operation type (addition and multiplication) and task complexity, defined not only based on surface criteria (number size), but also on individual participant's strategy ratings. Their findings do not support predictions of embodied numerical cognition theories which claim grounding of basic operation types in sensory and motor systems. Their fundings suggest that differences between procedural and fact-retrieval strategies in fronto-parietal and sensory-motor regions support the idea that verbal and sensory-motor derived concepts may play a role in general problem solving; and that using arithmetics in studying complex problem solving requires taking into account individual differences in skill, practice and strategy.

Verner [8] investigated cortical oxygen consumption on the verbally presented mental arithmetic task of addition, controlling for dominant strategy: retrieval or procedural. They claimed that the complexity of arithmetic tasks is mainly determined by the number of necessary additional cognitive processes, for example transferring carry digits. For simple arithmetic additions it is often enough to retrieve a fact from a memory, like 2+3=5. In order to perform an addition operation with large addends, one needs to use the retrieval strategy to get basic arithmetic facts, transfer a carry digit (perhaps several times), temporarily hold intermediate results in a memory and update results to get the answer, e.g. 23+19=20+10+9+3=30+10+2=42, or 23+19=23+7+12=30+12=42. For their study Verner defined three regions of interest (ROI) on each hemisphere: the inferior frontal gyri (IFG), the middle frontal gyri (MFG), and the superior frontal gyri (SFG). They found that, with respect to task difficulty, complex addition tasks, compared to simple addition task, led to higher oxygenation in all defined ROI except left IFG. Among all ROIs, IFG was more sensitive to the task itself and showed highest oxygenation compared to MFG and SFG.

While some studies have focused on effects of different operations used in mental arithmetic, others investigated the role of working memory in handling carries on multi-digit mental addition problems. For example, the study by [30] manipulated working memory load , the number of carries and the value of carries, in order to investigate the role of phonological and executive working-memory components in the carry operation in mental arithmetic. Their results with respect to working-memory load suggest that mainly the central executive is important in handling the number of carry operations as well as the value that has to be carried.

4 Mental calculations with different operand sizes

The goal of the first experiment was to test the assumptions that cognitive load increases with respect to increase in complexity of the mental calculation task, and that manipulating the presentation of the task has the predictable effect on the cognitive load. An additional goal was to investigate the feasibility of the experimental setup with simultaneous measurements of the hemodynamic brain activity and pupillary response and develop algorithms necessary for data preprocessing.

The mental arithmetic task was an addition with three levels of complexity, defined by the surface parameter (operand magnitude). The assumption was that single-digit addends whose sum is less than ten would require the retrieval strategy to be used, therefore less or no working memory load; double-digit and triple-digit addends would require the procedural strategy. Expected working memory load would be the smallest in easy tasks and largest in hard tasks. The control task was selected to be a non-arithmetic number identification task. Regions of interest for brain imaging, as in [8], are bilateral areas of the prefrontal cortex: IFG, MFG, SFG.

4.1 Materials and methods

4.1.1 Stimuli

A control task was designed to gather data on conditions, which include actions required to complete each task, but do not include mental calculation itself, hence it serves as baseline conditions for a mental calculation task. In the control task stimuli is single-, double- or triple-digit number with the addition operand on left or right side. Stimuli appeared in white color and vertical size of 3 cm on a gray background. Static stimuli stayed on the screen until presented number was correctly typed with the laptop keyboard. Dynamic stimuli appeared on the top of the screen and moved vertically towards the bottom of the screen with the speed of 1,5-3 cm/sec. Dynamic stimuli disappeared as soon as participant typed the answer correctly, or when stimuli reached the bottom of the screen. Speed of vertical movement of the dynamic stimuli was changed as a function of task difficulty. Examples of the control task: 5+, 13+, +721.

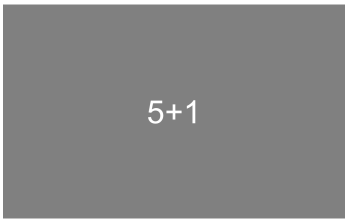
In the mental calculation task stimuli had same surface features as in the control task: an arithmetic expression of addition with two single-, double- or triple-digit addends. Single-digit addends correspond to the easy complexity level and their sum is < 10; double-digit addends correspond to the moderate complexity level and their sum is < 100; triple-digit addends correspond to hard complexity level and their sum is < 1000. Some examples of the mental calculation task are: 5+8, 14+26, 182+524. figure 1 illustrates stimuli of easy complexity, captured at different times.

4.1.2 Procedure

The study included two subjects, right-handed males. For each subject the study took 1.5 hours. The subjects received no compensation. The experimental session included standard procedures such as informative content of the participant, setting

(a) t=0s

(b) t=1.6s



(c) t=3.2s

(d) t=4.8s

(e) t=6.4s

Figure 1: Stimulus of the dynamic Presentation and easy Complexity, captured at different times on the screen.

up the equipment, performing four rounds of tasks, and receiving feedback from participant. Each subject's brain had to be scanned with MRI prior to the experiment in order to create an anatomical image. Subjects were wearing eye tracking glasses and optical imaging probe, attached to their forehead using flexible bandage. In addition, subjects underwent non-invasive continuous blood pressure monitoring using GE Healthcare B650 monitors (ECG and photoplethysmography were used to estimate blood pressure) to make sure that brain imaging data are not under global systemic change as a response to the task load. Stimulus were presented using HP EliteBook 8570w Mobile Workstation (laptop) with software PhyshoPy v.1.8.

Subjects received instructions before the task, performed the task and filled the task load index questionnaire after each round of the tasks. Control tasks (static, then dynamic) were presented in the first and second round, respectively; mental calculation tasks (static, then dynamic) were presented in third and fourth round, respectively. During each round 90 tasks were presented in 9 blocks of 10 tasks each. There were three blocks per each task complexity level (easy, moderate, fast), and the order of blocks was random. There were jitter resting time (5-15 seconds) between blocks. During the resting time, the participant was asked to look at the white fixation cross in the center on the screen.

Subjects were asked to solve a fixed number of mental arithmetic tasks. They were encouraged to solve tasks quickly, but were told that being correct is more important. The task on the control condition was to type the number, which was present on the screen, using the laptop keyboard. The task on mental calculation condition was to calculate the answer to an arithmetic expression, present on the screen, and type the answer using the laptop keyboard. It was emphasized that the answer should be retrieved mentally, i.e. without using finger counting, verbalization or any external device. Tasks on the static condition were presented in the middle of the screen. The participant had no time limitation to answer. Next task was shown as soon as correct answer to the current task was typed. Tasks on the dynamic condition were presented in a dynamic way: each task appeared on the top of the screen and was moving vertically to the bottom with a constant speed. Participant had an implicit time limitation to answer: the answer should be entered before the expression reached the bottom of the screen. The next task was shown as soon as the correct answer to the current task was typed. If no correct answer was given, next task was shown after current task reached the bottom of the screen.

On the mental calculation condition, complexity of the task (referred to as task complexity or intrinsic complexity) varied across three levels: easy, moderate, and hard. Each mental calculation tasks was expressed as an addition of two operands. On dynamic condition speed of the movement (referred to as presentation complexity or extraneous complexity) varied across three levels in the following way: slow speed level was assigned to the tasks of hard complexity, moderate speed level was assigned to the tasks of moderate complexity, fast speed level was assigned to the tasks of easy complexity.

4.1.3 Analysis

The following data were collected during the experiment: automatically logged behavioral data (responses, response times, accuracy of the responses), behavioral data in form of questionnaire (post-task subjective load assessment), physiological data (pupil size measurements, brain hemodynamic measurements), video material (view of a subject throughout the task).

Answers to the equations and response times were collected by the stimulus presentation software PsychoPy with an customized Python script, running on a laptop HP EliteBook 8570w with OS Windows 10. These data included the following:

- Correct answer. Answer is a sequence of (numeric) key presses, which is expected by the software as result of addition of two presented numbers. E.g., if the stimulus is 12+13, then correct answer is 25, which should be typed as sequence 2,5. There can be at most one correct answer per task.

- Incorrect answer. Answer is incorrect if the typed sequence does not match the sequence of characters in a correct answer. E.g., if the correct answer is 25 and entered keys are 4,5,2,5, then two incorrect answers were given (4,5) and one correct answer (2,5). There can be many missing answers per task.

- Missing answer. On the dynamic conditions, it participant fails to type the answer while stimuli is on the screen, answer counts as missing. There can be at most one missing answer per task.

- Time to first answer. Time between stimulus presentation and first entered numeric key (in seconds).

- Time to correct answer. Time between stimulus presentation and last entered numeric key of the correct answer (in seconds).

VALIDATION OF THE TASK DIFFICULTY

Subjective task load index questionnaire and response time comparison were used to validate the manipulation of the task difficulty.

Task load

To ensure that tasks of different complexity had induced distinguishable difficulty in participants, post-round NASA TLX (Task-Load Index) [31] subjective workload assignment and task performance measures were used. Raw TLX scores for each of accessed questions, are presented in figure 2, aggregated and normalized TXL scores illustrated in figure 3. The normalized scores are given at the table 1.

Sum of scores for features were compared across conditions for both subjects. As the figure 3 illustrates, manipulation of task complexity was successful only for one of subjects, where increase in Complexity (easy, moderate, hard) and change of Presentation (static, dynamic) resulted in increase in normalized task demand score. For another subject on static condition mental workload for easy tasks reported to be nearly same as for hard tasks, and mental workload for moderate tasks was the

Table 1: Aggregated and normalized TLX scores by task complexity for each subject and presentation separately.

Subject	Presentation	Task complexity		
		easy	modr	hard
J1	static	0.075	0.317	0.417
	dynamic	0.133	0.400	0.633
S1	static	0.333	0.467	0.367
	dynamic	0.517	0.450	0.467

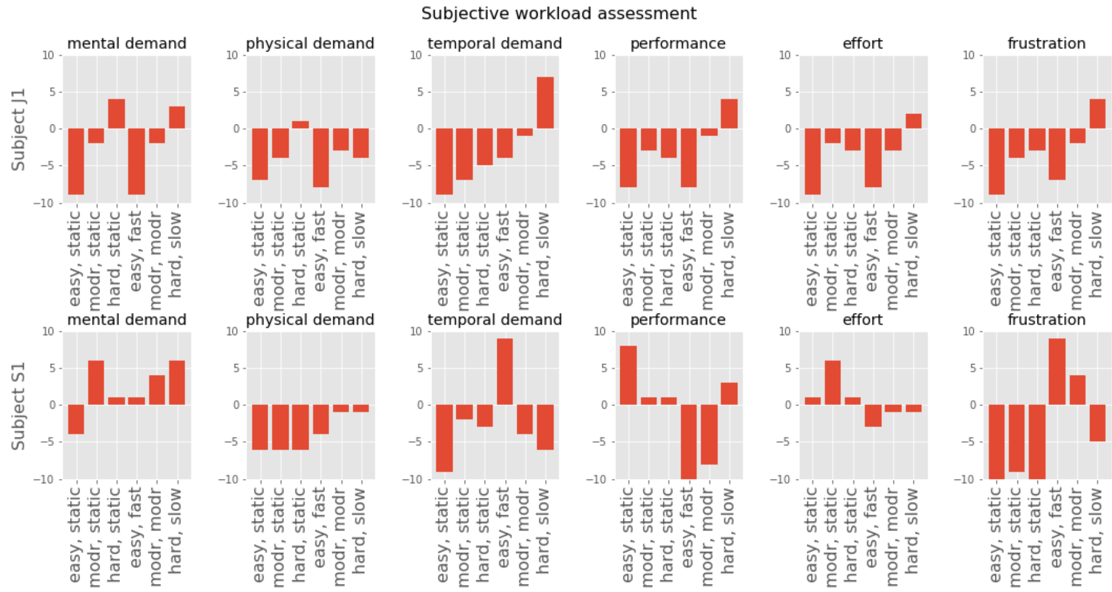


Figure 2: Raw TLX scores of mental calculation tasks, grouped by subject and condition and plotted for subjects and scored quantities separately. According to the NASA TLX questionnaire, for all categories except the Performance negative value corresponds with the quantity described as Very Low, positive value corresponds with the quantity described as Very High. For Performance the negative value refers to Perfect, the positive value refers to Failure. It is likely that the subject S1 had rated the Performance on easy static task close to Failure mistakenly because of such inconsistency.

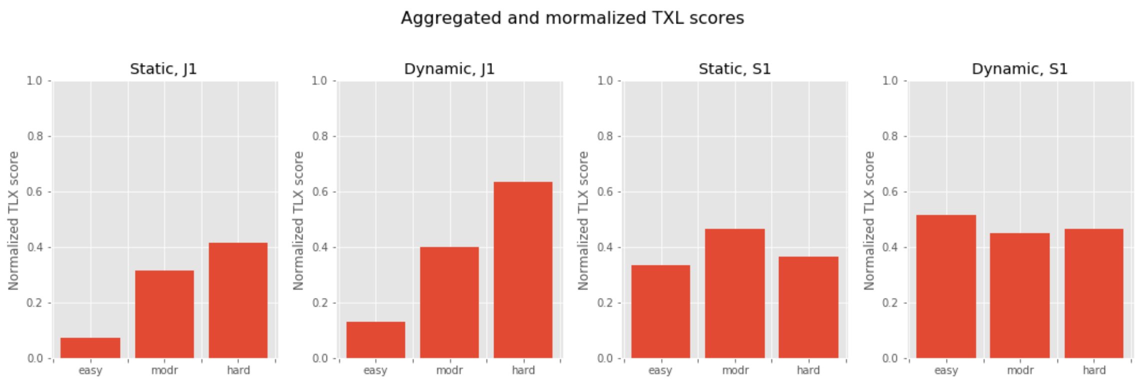


Figure 3: Normalized TLX scores of mental calculation tasks, grouped by task complexity and plotted for subjects and presentation separately.

largest; on dynamic condition mental workload for moderate tasks reported to be nearly same as for hard tasks, and mental workload for easy tasks was the largest.

Response times

Response time is the time to enter last numeric key of the correct answer in Calculation condition minus average time to enter last numeric key of the correct answer in Control condition (separately for each Subject, Complexity and Presentation). In other words, response time is the time between stimuli appears on the screen and last key of the correct response is typed, minus time required to actually type the characters. Incorrect responses (4% of all responses) were not considered, response times more than two standard deviations were removed from analysis as outliers. Average response times for both subjects presented in Table 2. The figure 4 illustrates the results.

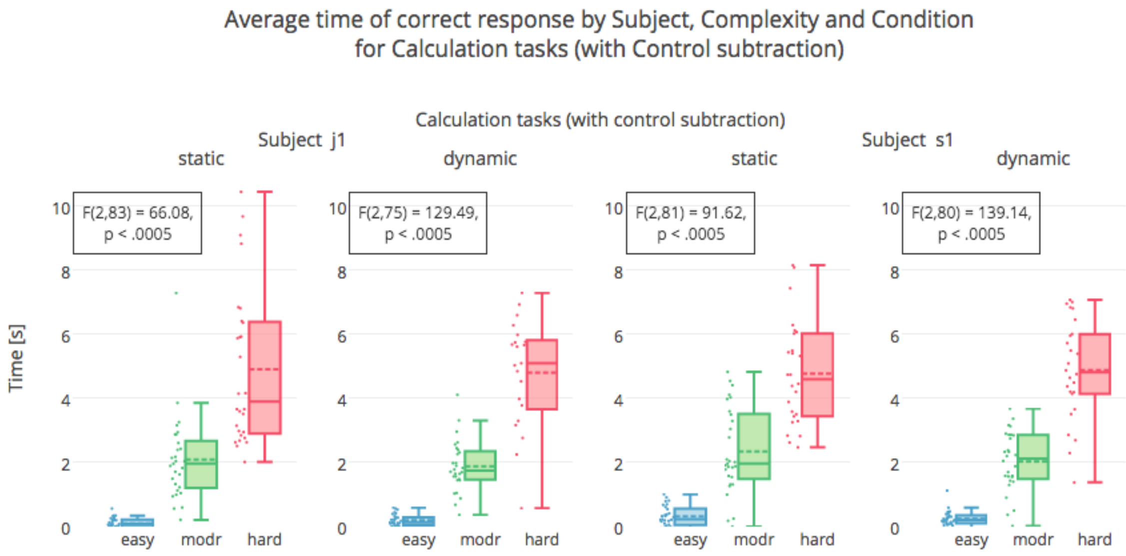


Figure 4: Response times of the subjects (with the subtraction of last key press of the correct answer on the control task), grouped by Condition and Complexity

The grand mean for the response time was 2.27 seconds. As evident in the means, on the static condition for subject J1 moderate tasks took 1.7s longer to answer than easy tasks, hard tasks took 2.8s longer than moderate tasks. The difference was statistically significant ($F(2,83)=66.08$, $p<.0005$). On the dynamic condition moderate tasks took 1.7s longer to answer than easy tasks, hard tasks took 2.8s longer than moderate tasks. The difference was statistically significant ($F(2,75)=129.49$, $p<.0005$).

For subject S1 on static condition moderate tasks took 2.0s longer to answer than easy tasks, hard tasks took 2.4s longer than moderate tasks. The difference was statistically significant ($F(2,81)=91.62$, $p<.0005$). On the dynamic condition moderate tasks took 1.8s longer to answer than easy tasks, hard tasks took 2.9s longer than moderate tasks. The difference was statistically significant ($F(2,80)=139.14$, $p<.0005$).

For both subjects there was no significant difference between static and dynamic condition (J1: $F(1,158)=0.1$,ns; S1: $F(1,161)=0.3$, ns), nor between Complexity and Presentation (J1: $F(2,158)=0.2$,ns; S1: $F(2,161)=0.5$, ns).

Table 2: Average response times (in seconds) by task complexity for each subject and presentation separately.

Subject	Presentation	Task complexity		
		easy	modr	hard
J1	static	0.096	2.094	4.913
	dynamic	0.197	1.884	4.809
S1	static	0.319	2.347	4.778
	dynamic	0.245	2.034	4.887

PUPILLARY RESPONSE ANALYSIS

Complex visual information, e.g. light conditions and luminance, affects the pupil size. Cognitive tasks, such as mental calculation or working memory load also lead to very robust increases in pupil size. Similarly, pupil size responds to physical effort, violation of predictions, shifts in the exploration/exploitation tradeoffs [32], etc. A common issue in analysis of the pupillary time series is that data are non-linear, non-stationary and might contain low-frequency drifts along with high-frequency noise. Drifts, tremors and non-spherical eye shape might also introduce noise into pupil size signal.

In block design experiments, samples of pupillary response can be averaged over a time window and then compared across conditions. In the experiments with fast-paced events it can be hard to analyze pupil size with respect to each event separately because of the superposition of the delayed pupil responses. Design of this experiment is such that data of each round contain ten unevenly distributed events within each of nine blocks, thus traditional design window averaging approach is complicated.

Different methodological approaches might be used for analyzing pupil size. Baseline subtraction (subtracting the mean of the pupil size data taken from a portion of time preceding the event onset) is the most common method. This method requires many trial repetitions in order for the noise to cancel out and does not deal with the slope of the signal at the event onset. Other works, e.g. [6] argue that some of high-frequency noise can be removed from data by smoothing it with low-pass filter, where cutoff frequency might be determined using correlation between left and right pupil sizes at different frequencies. Hoeks and Levelt [33] have proposed the use of a deconvolution technique to analyze the pupil size data. In their method, the pupil size data is deconvolved with a canonical impulse response. The approach is analogous to the convolution of design matrix and hemodynamic response function in fMRI experiments. Issues with this approach include high between-subject variability in the shape of the impulse response and the delay between the event onset and the beginning of the response. Finally, [32] proposed a system identification framework to optimize the analysis of pupillometric data, where low-frequency noise is modeled with an autoregressive model, and the pupillary response to specific events is modeled by adding exogenous inputs.

Data preprocessing

In this work the pupil measurements were collected while subjects were performing tasks. Participant's left and right eye pupil diameters were measured using an Eye Tracking Glasses Natural Gaze (SensoMotoric Instruments, SMI, Teltow, Germany) at a sampling rate of 30 Hz, and data recorded using the iView X Software (SMI, Teltow, Germany). Pupillometry data were preprocessed using a custom made Python script to remove artifacts in the time series related to eye blinks and saccades. Data points, detected as Blinks by the iView X Software and data points with physiologically unlikely pupil sizes (smaller than 2 mm or larger than 7 mm) together with the neighboring data points (the preceding and following 30 ms) were removed. Resulting gaps in the data were filled with linear interpolation. In both subjects data from the left pupil were much more noisy compared to data from the right pupil.

Table 3: The average pupil sizes for complexity levels

Subject	Condition	Presentation	Task complexity		
			easy	modr	hard
J1	baseline	dynamic	3.43	3.31	3.32
		static	3.71	3.66	3.75
	calculation	dynamic	3.31	3.31	3.42
		static	3.32	3.39	3.46
S1	baseline	dynamic	3.77	3.83	4.00
		static	4.23	3.99	4.02
	calculation	dynamic	4.04	3.94	3.95
		static	3.79	3.66	3.76

This happened probably due to interference of the cords of the optical probe with the field of view. Therefore, only data from right pupil were used in the analysis. The example of the raw and cleaned data at all preprocessing steps illustrated by figure 5.

Trial samples at first 60ms from the onset were considered a baseline and subtracted from the rest of the trial data. The preprocessed time series were averaged for each task complexity and presentation and plotted on the timescale from the onset of the stimuli to the beginning of the response. The resulting time series of each round for both subjects are illustrated by figures 6 and 7.

The average pupil sizes for complexity levels are given in Table 3. The two-way within-subject ANOVA was conducted to compare the effect of task complexity, presentation and interaction between task complexity and presentation on pupil size change. Differences in pupil size change for both subjects and both conditions (baseline, calculation) were tested independently of each other with two-way analysis of variance (ANOVA) for dependence on factors Complexity (easy, moderate, hard) and Presentation (static or dynamic). Statistical significance values are reported in Table 4.

HEMODYNAMIC BRAIN ACTIVITY ANALYSIS

Hemodynamic traces of brain activity were measured using intensity modulated frequency-domain optical tomography system [34] with 16 channels. A flexible probe of optical fibers, fiber bundles and prisms was used. The approximate dimensions of the probe are 80mm x 130mm x 6mm. The probe was located on the subject's forehead in a way to cover prefrontal cortex areas, as shown on Image 8. The fiber bundles were moved to place the eye tracking glasses, as shown on Image 9.

The measured source - detector distances ranged from 8 mm to 98 mm, however, in order to avoid contribution from possible light leaks on the measured signal, source-detector pairs were limited to 20-35 mm distance in this work. The selected source - detector pairs were labeled according to the brain region they approximately cover, and three regions were of primary interest: superior frontal gyri (SFG), middle frontal gyri (MFG), and inferior frontal gyri (IFG). The study by [8], among others, have shown that, with respect to task difficulty, more complex addition tasks lead

Table 4: Statistical significance values for factors Task complexity and Presentation

Subject	Condition	Effects of factors	Significance F	df	p(>F)
J1	baseline	Task complexity (easy, moderate, hard)	69	2	<.005
		Presentation (static, dynamic)	3783	1	0
		Task complexity : Presentation	50	2	<.005
	calculation	Task complexity (easy, moderate, hard)	696	2	<.005
		Presentation (static, dynamic)	358	1	<.005
		Task complexity : Presentation	29	2	<.005
S1	baseline	Task complexity (easy, moderate, hard)	149	2	<.005
		Presentation (static, dynamic)	1133	1	<.005
		Task complexity : Presentation	504	2	<.005
	calculation	Task complexity (easy, moderate, hard)	345	2	<.005
		Presentation (static, dynamic)	6002	1	0
		Task complexity : Presentation	107	2	<.005

to higher oxygenation in all defined ROI except in the left IFG compared to simple addition tasks. The Image 10 indicates position of regions of interest in the cortex.

The signal drift caused by contact variations, sweating and instrument drift was removed by subtracting a low-pass filtered version (cutoff frequency 0.007Hz) of the amplitude signal from the time course. The example of such preprocessing illustrates by figure 11. Oxygenated, deoxygenated and total hemoglobin concentration changes in the brain were estimated based on changes in measured data.

4.2 Results

The analysis of variances indicated that effects of Task complexity, Presentation and interaction between Task complexity and Presentation are all significant at .005 level. The comparisons of the average pupil sizes indicate that for subject J1 on static calculation condition the pupil diameter increased with the increase of task complexity, $\Delta(easy, moderate)$ = 0.07mm, $\Delta(moderate, hard)$ = 0.07mm. On the dynamic calculation for same subject pupil size on hard tasks were on average 0.11 mm smaller than on easy and moderate tasks. For baseline condition, moderate tasks were associated with the smallest pupil size, and easy tasks were associated with the largest pupil size.

Similar comparison of average pupil size for subject S1 revealed that for static calculation condition pupil size is smallest on moderate tasks, and largest on easy tasks: $\Delta(easy, moderate)$ = -0.07mm, $\Delta(moderate, hard)$ = 0.1mm. On the dynamic calculation for same subject pupil size on hard tasks were on average 0.09 mm smaller than on easy and moderate tasks. For baseline static condition, moderate tasks were associated with the smallest pupil size, easy tasks were associated with the largest pupil size. For baseline dynamic condition, smallest average pupil size was observed at easy condition, and largest at hard condition.

Interestingly, in some cases average pupil size on calculation condition was smaller then on baseline counterpart (e.g. all levels of J1, static; easy level of S1, static). This contradicts the expectation, but can likely be explained by the flaws in the experiment design.

The average value of the measurement alone is not enough to draw conclusions about the time series data, because it ignores the time course of the data. Closer look at the figures 6 and 7 reveals that for on baseline condition pupil size does rarely increases more than 0.1mm, where on calculation condition increase over time is up to 0.4mm. The time course of the response for subject S1 on static calculation condition clearly shows the peak difference in pupil sizes changes at approx. 3.5 seconds from the onset of the stimuli, with $\Delta(moderate, hard)$ approx. 0.4mm.

As for hemodynamic brain activity, each source-detector pair have been labeled as one of regions of interest: SFG, MFG, IFG. Differences in total hemoglobin concentration changes were tested for dependence on factors Complexity (levels easy, moderate, hard) and Time for each or ROIs separately. The two-way ANOVA revealed significant effect of Complexity on hemoglobin concentration change in all regions (SFG: $F(3{,}1114126)=1174$, $p=0$; MFG: $F(3{,}165460)=11.50$, $p<.005$; IFG: $F(3{,}1036909)=1196$, $p=0$).

The results indicate that differences in task Complexity levels correspond to variations in pupil size and total hemoglobin concentration change in the brain. The analysis of pupillary responses has shown that, although sensitive to variations in the task complexity in some cases, pupil size was much more dependent on the luminance of the stimuli.

4.3 Discussion

In accordance with the cognitive load theory, the cognitive load varied in two aspects: intrinsic and extraneous. Intrinsic component was expected to change according to three levels of difficulty (easy, moderate and hard) of the addition task. Extraneous component was introduced by the presentation of the task (speed of the movement: slow, moderate, fast) with the purpose to mimic the complexity added by instructional design in particular, or user interface design in general. However, the experimental design had some limitations.

First, although the block order was randomized across the participants, experimental rounds have followed the fixed order: control static, control dynamic, calculation static, calculation dynamic. In addition to that, the three levels of the dynamic presentation (fast, moderate, easy) were confounded with the three levels of the task complexity, such that dynamic tasks of the easy level were always fast, dynamic tasks of the hard level were always slow.

Secondly, each block had a fixed number of stimuli, ten. In large part of the tasks this lead to an answer time of approximately one second, although the initial contraction of the pupil due to a light reflex happened after approximately 1 - 1.5 seconds after the onset of the stimuli. There were no between-stimulus resting time, therefore it was not possible to extract the event-related pupillary response for such fast-paced stimuli.

Lastly, the luminance of the stimuli was not controlled, and levels of the task difficulty might have been confounded with the luminance: stimulus of easy level occurred as three white characters on a gray screen (including the plus sign), stimulus of moderate level occurred as five characters, and stimulus of hard level had seven characters.

The second experiment was designed in order to eliminate these drawbacks and manipulate the conditions more carefully.

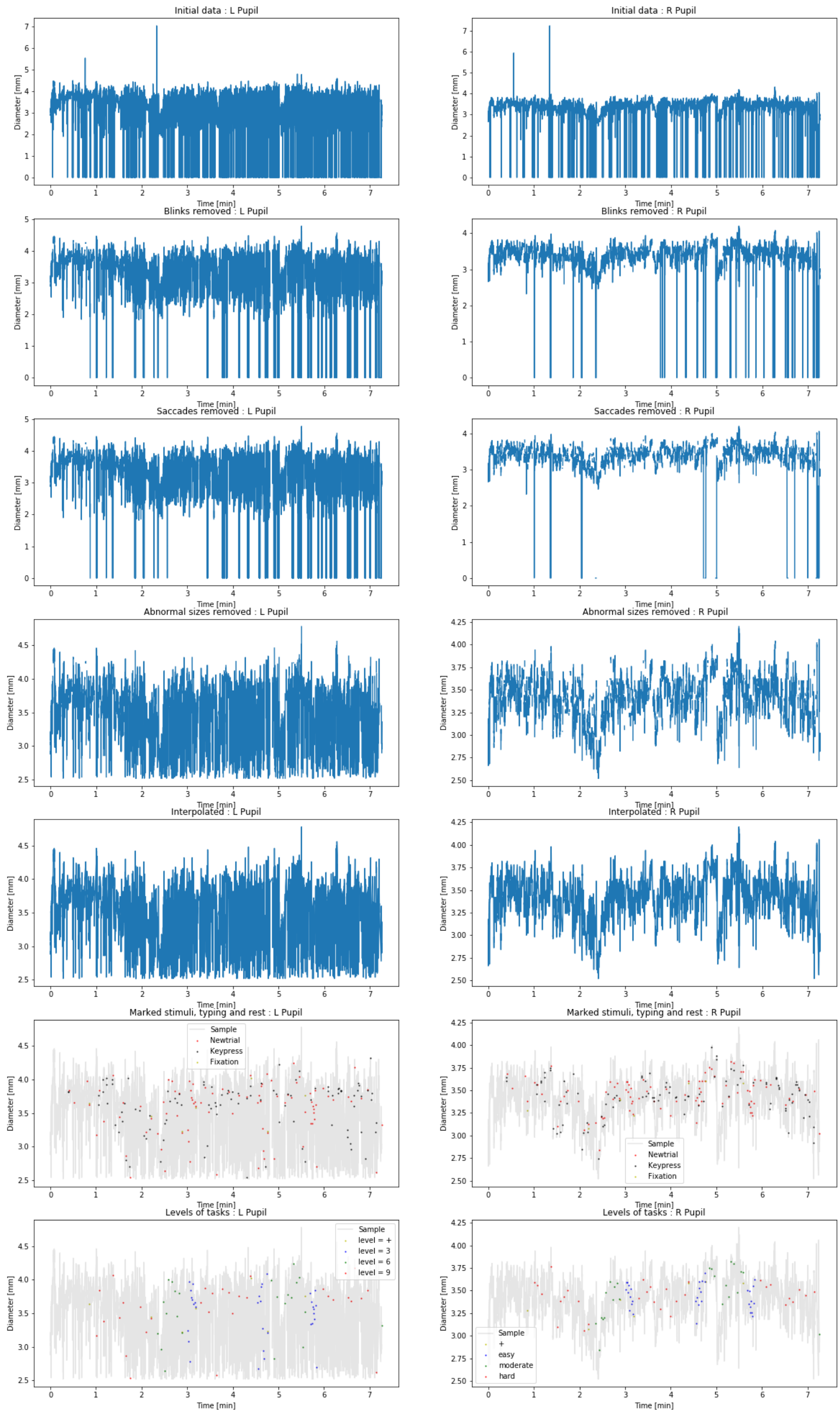
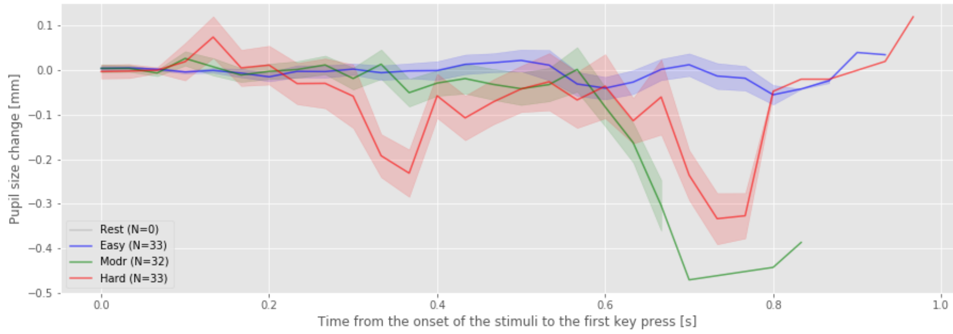
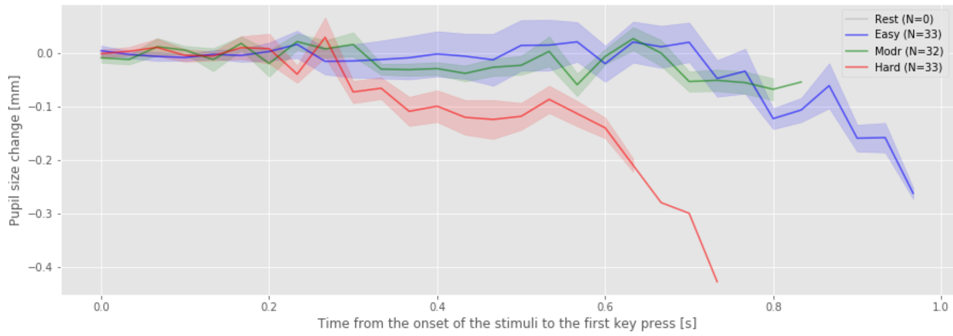


Figure 5: The illustration of the preprocessing steps for one of the subjects. Left column illustrates pupil size time series for left pupil, right column - for right pupil. Initial data plotted at the first row; then data after removal of blinks and saccades; fourth row contains data after removal of abnormal pupil sizes; in fifth row interpolated data plotted. Row six contains data with markers of trial onset, key response onset, and a fixation onset. Finally, last row illustrates onsets of stimuli of different task complexity.



(a) Static baseline condition

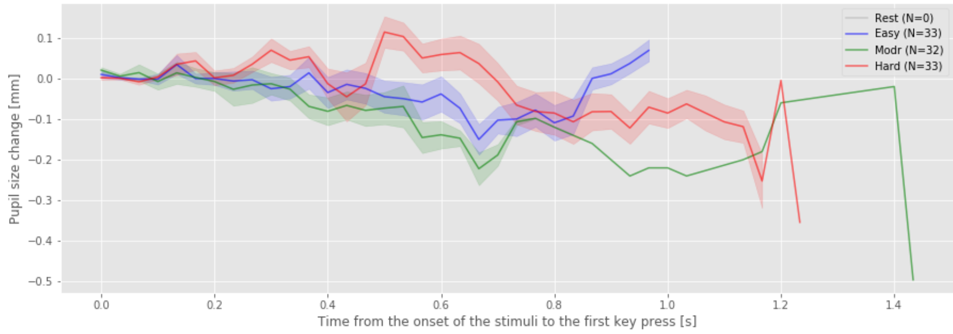


(b) Dynamic baseline condition

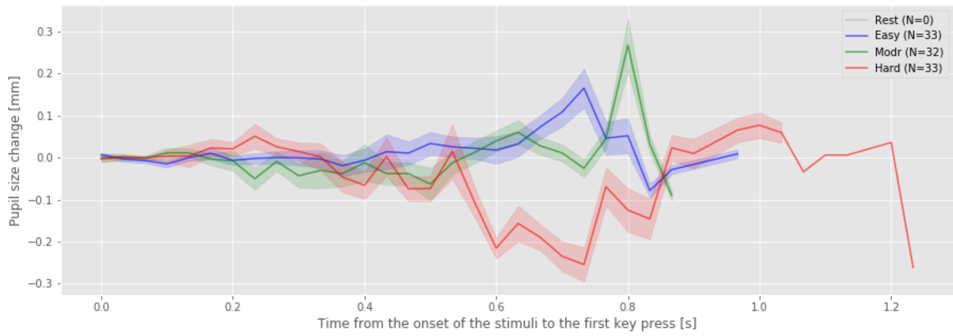
(c) Static calculation condition

(d) Dynamic calculation condition

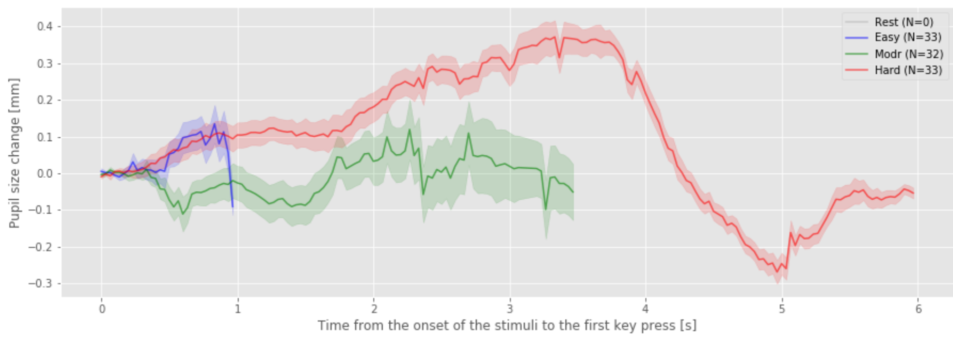
Figure 6: The preprocessed and task-separated pupillary size data for for all experimental rounds of subject J1. The pupil diameter is plotted from the onset of the stimuli to the first key press.



(a) Static baseline condition



(b) Dynamic baseline condition



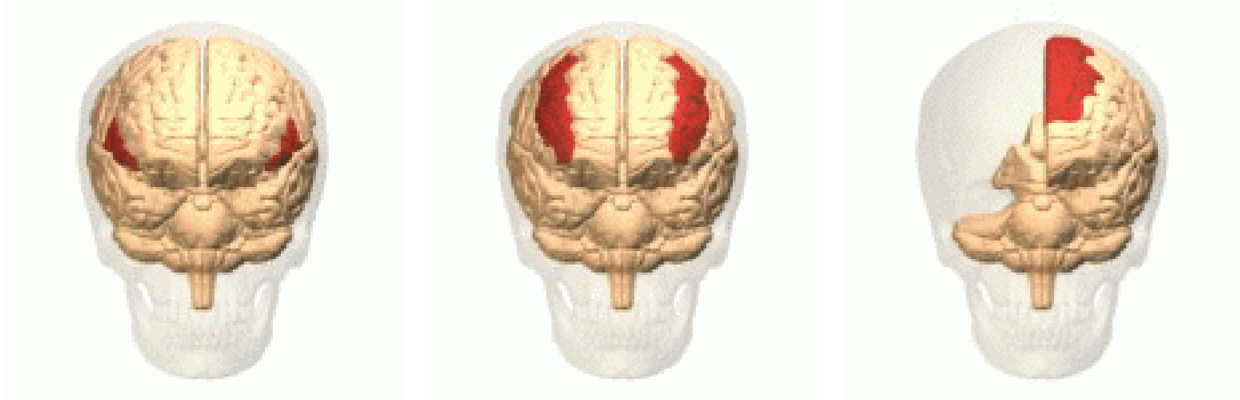
(c) Static calculation condition

(d) Dynamic calculation condition

Figure 7: The preprocessed and task-separated pupillary size data for for all experimental rounds of subject S1. The pupil diameter is plotted from the onset of the stimuli to the first key press.

Figure 8: Position of the neuroimaging probe on the forehead of a subject.

Figure 9: Position of the neuroimaging probe and the eye tracking glasses on the forehead of a subject.



(a) Inferior frontal gyrus

(b) Middle frontal gyrus

(c) Superior frontal gyrus

Figure 10: Location of the regions of interest in the brain cortex.

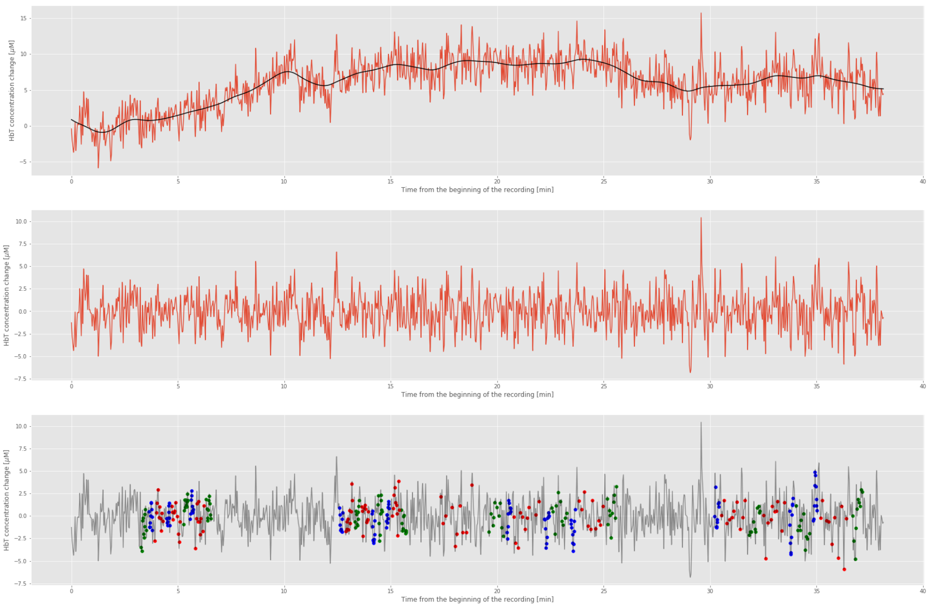


Figure 11: The example of hemodynamic data preprocessing. The top row shows the raw data for one of source detector pairs. The middle row is the data with subtracted low-frequency drift. The bottom row illustrates the onsets of the stimuli (blue - easy tasks, green - moderate tasks, red - hard tasks).

5 Mental calculations with same operand sizes

The experiment was designed with the idea of decreasing the signal superposition due to fast-paced responses, controlling the luminance and eliminating confounding of the task presentation and the complexity. The previous studies, e.g. [30], have shown that the number of carries in mental addition tasks determines the problem difficulty and that it is mediated by the executive subsystem of the working memory. The new set of tasks has been constructed in the following way: each task is a three-digit addition task with the complexity level determined by the number of required carry operations.

5.1 Materials and methods

5.1.1 Stimuli

Tasks of level 0 require no carry operations, e.g. 123+254. Tasks of level 1 require one carry operation, which always occurs in the ones, e.g. 245+318 requires a carry from 5+8. Tasks of level 2 require two carry operations, which always occur in tens and ones, e.g. 356+479 requires a carry from 6+9 and a carry from 5+7. The control tasks were presented as a three-digit number plus zeros, e.g. 549+000, or 000+168.

In each trial, a stimuli appeared in white color and vertical size of 30mm. In static condition the stimuli appeared in the middle of the screen, in slow dynamic condition the stimuli was moving from the top to the bottom with the speed of 1.15cm/s, in fast dynamic condition the speed of moving was 2.3cm/s.

5.1.2 Procedure

The study included four subjects, three males and a female, with the mean age 29 years. The study took 2 hours for each subject, 30 minutes of preparation and setting up, and 90 minutes of the recording.

The stimuli were presented in one of three conditions: static, slow dynamic, and fast dynamic. Levels of complexity (including control) and presentation were mixed randomly. Each stimuli was followed by 3-5 seconds of the resting time. Every 150 seconds there was a longer rest of 25-35 seconds, in total 90 minutes on task performance from each participant has been recorded.

Subjects were wearing eye tracking glasses and optical imaging probe, attached to their forehead using flexible bandage, they underwent non-invasive continuous blood pressure monitors, as in the previous experiment. Stimulus were presented using HP EliteBook 8570w Mobile Workstation (laptop) with software PhyshoPy v.1.8.

5.1.3 Analysis

VALIDATION OF TASK DIFFICULTY

Task difficulty was validated using response times for the correct trials. The grand mean response time was 5.98 seconds. The average correct response time for tasks of complexity 0 was 4.95 seconds, for tasks of complexity 1 was 6.26 seconds, for tasks of complexity 2 - 7.05 seconds. The average correct response times for different presentation levels were 5.93 seconds for static presentation, 6.24 seconds for slow dynamic presentation, and 5.75 seconds for fast dynamic presentation. The effects of Complexity, Presentation and their interaction were tested with 2x2 ANOVA. The analysis of variances revealed a statistically significant effect of Complexity on correct response time ($F(2{,}1239)=70.38$, $p<.005$), a statistically significant effect of Presentation on correct response time ($F(2{,}1239)=7.87$, $p<.005$), and no statistically significant interaction between Complexity and Presentation ($F(4{,}1239)=1.49$, $p>.05$).

PUPILLARY RESPONSE ANALYSIS

The instrumentation, measurement and preprocessing of the pupillary data were similar to the process used in the first experiment. Trial samples at first 60ms from the onset were considered a baseline and subtracted from the rest of the trial data. The preprocessed time series were averaged for each task complexity and presentation and plotted on the timescale from the onset of the stimuli to the beginning of the response. The panels with pupillary time series for each participant are illustrated by figures 12, 13, 14, 15.

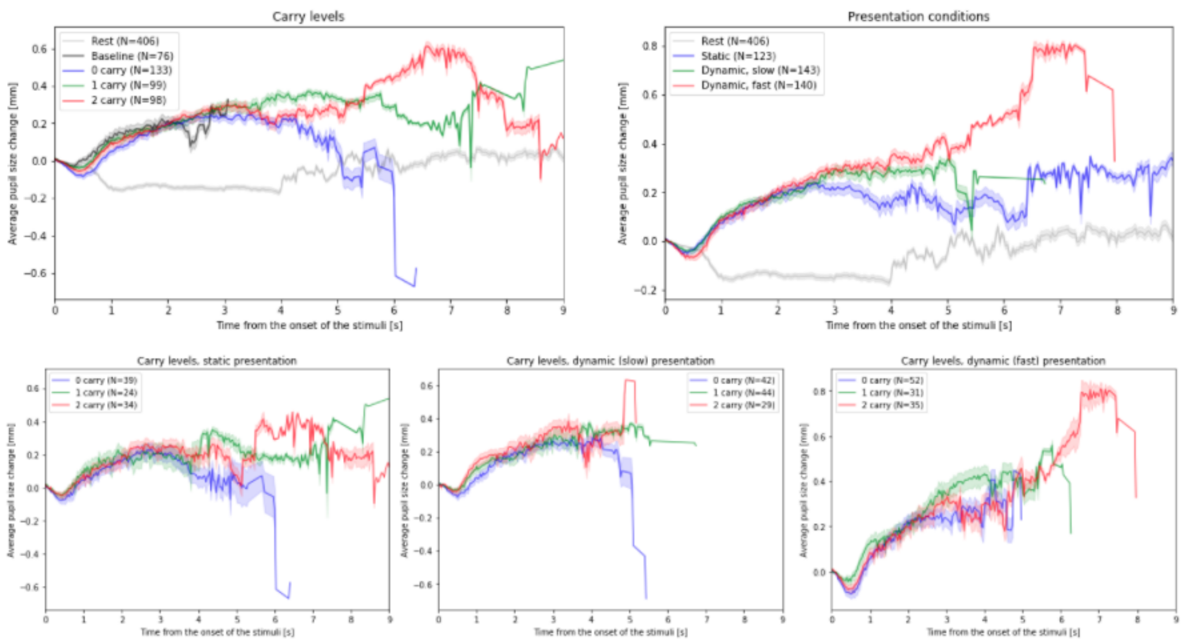


Figure 12: Comparison of pupil time series from onset of the stimuli. Top row of each panel shows pupil time course separately for each Complexity level (left) and separately for each Presentation level (right). Three plots in the middle row illustrate pupil time series for each combination of Complexity and Presentation. Subject S2.

On average, the answer of participant on the longest (static) trials was given at time 9 seconds from the stimuli presentation, therefore the data from the time interval [0,9] seconds was used for further analysis and statistical testing.

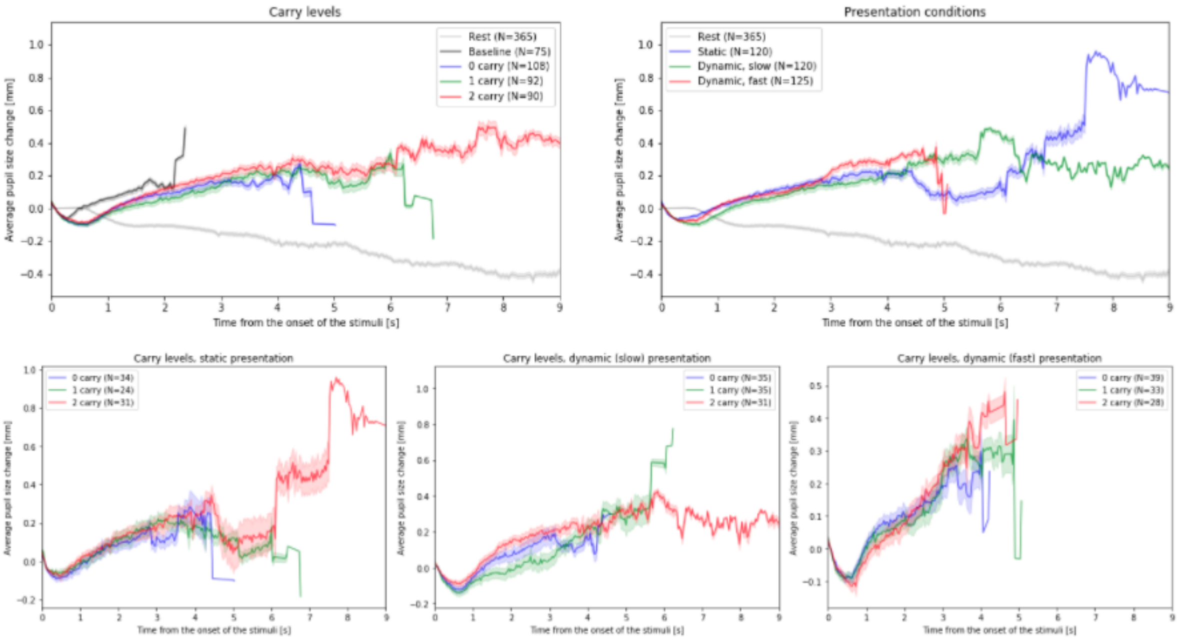


Figure 13: Comparison of pupil time series from onset of the stimuli. Top row of each panel shows pupil time course separately for each Complexity level (left) and separately for each Presentation level (right). Three plots in the middle row illustrate pupil time series for each combination of Complexity and Presentation. Subject J2.

HEMODYNAMIC BRAIN ACTIVITY ANALYSIS

The measurement probe was placed differently in order to cover a larger part of the right prefrontal cortex and to avoid introducing noise to pupil size measurements. The approximate placement of the probe is illustrated by figure 16.

The hemodynamic brain activity data was collected simultaneously with pupillary measurements. For one of the subjects, S2, the data was lost due to a technical error. Data of the rest of the subject were preprocessed similarly with the previous experiment. The low-frequency drifts were removed by subtracting the the low-pass filtered data (cutoff frequency 0.007Hz) from the signal. Data of each trial was baseline-corrected by subtracting two samples prior each stimuli onset (one second). The resulting data were averaged across complexity and presentation levels, and interval of ten seconds from the onset was used for statistical analysis.

Out of 210 source-detector pairs the expected relation between complexity levels was detected at, on average, 92 source - detector pairs, 18 source - detector pairs were same between three subjects. Most of these pairs had source and detector covering superior frontal gyrus area.

Table 5: Average change in pupil size for each participant by Presentation and Complexity

Subject	Presentation	Complexity				
		0 carry	1 carry	2 carry	baseline	fixation
	static	0.095	0.124	0.135	0.113	-0.119
S2	dynamic (slow)	0.119	0.151	0.170	0.099	-0.119
	dynamic (fast)	0.118	0.187	0.159	0.060	-0.119
J2	static	0.015	0.085	0.093	0.035	-0.107
	dynamic (slow)	0.027	-0.007	0.092	0.031	-0.107
	dynamic (fast)	0.045	0.047	0.051	0.035	-0.107
K2	static	-0.216	-0.124	-0.247	0.056	-0.078
	dynamic (slow)	-0.118	-0.081	-0.007	0.159	-0.078
	dynamic (fast)	-0.05	-0.042	-0.054	0.163	-0.078
R2	static	0.018	0.065	0.140	-0.031	-0.099
	dynamic (slow)	0.210	0.229	0.239	0.023	-0.099
	dynamic (fast)	0.235	0.279	0.245	0.042	-0.099

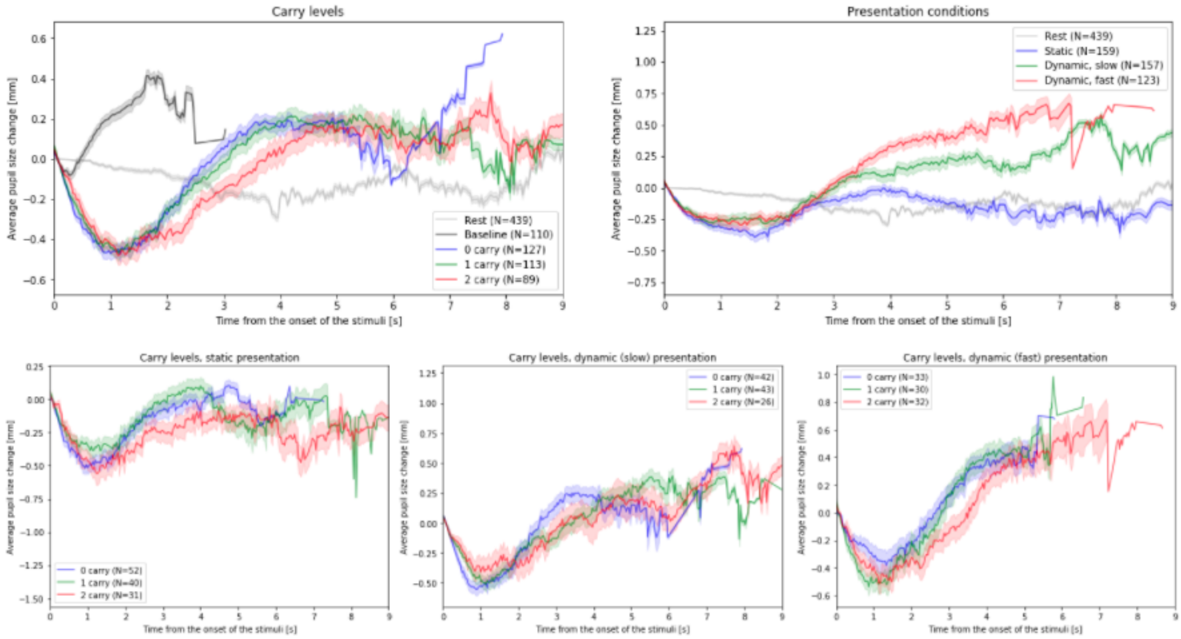


Figure 14: Comparison of pupil time series from onset of the stimuli. Top row of each panel shows pupil time course separately for each Complexity level (left) and separately for each Presentation level (right). Three plots in the middle row illustrate pupil time series for each combination of Complexity and Presentation. Subject K2.

5.2 Results

The average pupil sizes for complexity levels are given in Table 5. The three-way within-subject ANOVA was conducted to compare the effect of task complexity, presentation, and interaction between task complexity and presentation on pupil size change. An additional factor of onset time was user in order to allow comparison on non-independent samples from a time series. Differences in pupil size change for all subjects and conditions (baseline, calculation) were tested for dependence on factors Complexity (carry0, carry1, carry2), Presentation (static, dynamic (slow), dynamic (fast)), Time, and interaction between Complexity and Presentation. The effect of complexity level on pupil size change was significant ($F(4,651883)=3.88$, $p<.005$), the effect of presentation on pupil size change was significant ($F(3,651883)=5.17$, $p<.005$), the interaction between complexity level and presentation was not significant ($F(12,651883)=1.64$, $p>.05$).

The results of the pupil data analysis indicate that, with control for luminance, the pupil size can be a reliable indicator of the cognitive load. Trace differences between carry levels on all presentation conditions are illustrated by figure 17. In all cases, there was a large positive difference most of the time in pupil size between task and fixation, between task and baseline, and between most and least complex tasks (carry levels 2 and 0). A smaller positive difference was observed in most of the cases between tasks of carry levels 2 and 1, 1 and 0. As for the presentation

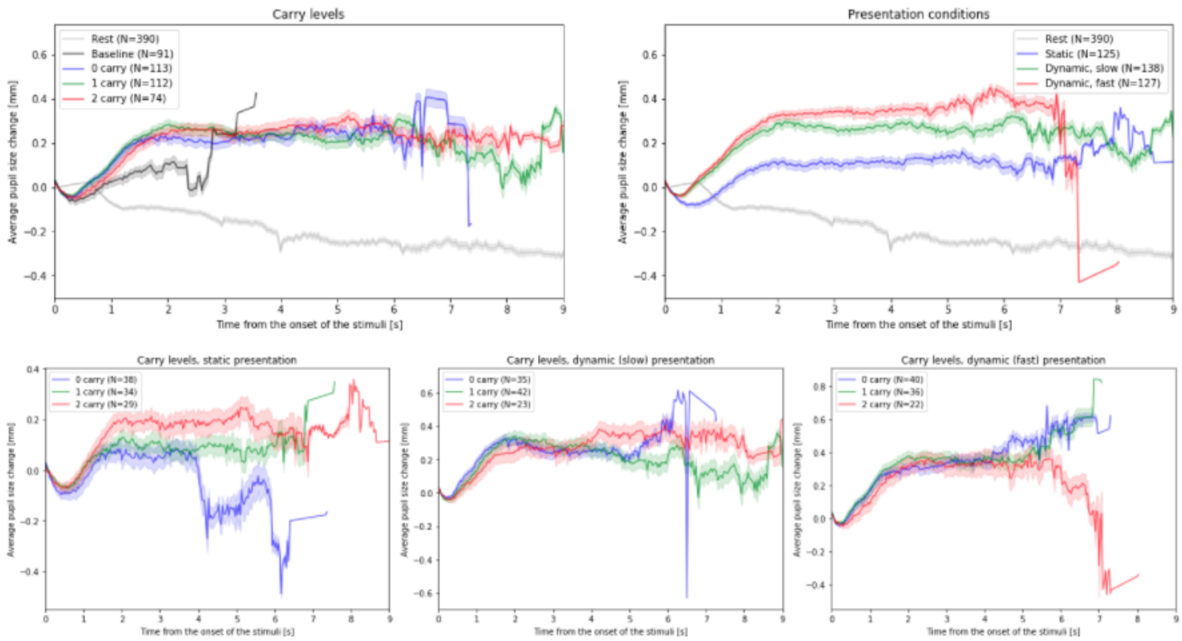


Figure 15: Comparison of pupil time series from onset of the stimuli. Top row of each panel shows pupil time course separately for each Complexity level (left) and separately for each Presentation level (right). Three plots in the middle row illustrate pupil time series for each combination of Complexity and Presentation. Subject R2.

difficulty, pupil size change was consistent with increasing presentation difficulty only when data from different complexity levels pulled together. When controlling for the complexity levels, there is no consistency between change in pupil size and change in the presentation complexity.

During hemodynamic activity analysis, each source - detector pair have been labeled as one of regions of interest: SFG, MFG, IFG. Differences in total hemoglobin concentration changes were tested for dependence on factors Complexity (levels 0, 1, 2), Presentation (static, dynamic (slow), dynamic (fast)), Time, and interaction between Complexity and Presentation for each or ROIs separately. The three-way ANOVA revealed significant effect of Complexity on hemoglobin concentration change in all regions (SFG: $F(5,10724533)=3036$, $p=0$; MFG: $F(5,1787408)=1245$, $p=0$; IFG: $F(5,2502378)=167$, $p<.005$), no significant effect of Presentation on hemoglobin concentration change (SFG: $F(2,10724533)<1$, ns; MFG: $F(5,1787408)<1$, ns; IFG: $F(5,2502378)<1$, ns), and significant interaction between Complexity and Presentation (SFG: $F(10,10724533)=99.09$, $p=<.005$; MFG: $F(10,1787408)=12.06$, $p<.005$; IFG: $F(10,2502378)=18$, $p<.005$).

The results indicate that differences in task Complexity levels correspond to variations in pupil size and total hemoglobin concentration change in the brain. For the Presentation, only pupil size changes were sensitive to the different levels of that factor. In addition, even though response times did not clearly indicated complexity of the task presentation, pupil size reflected different complexity levels.

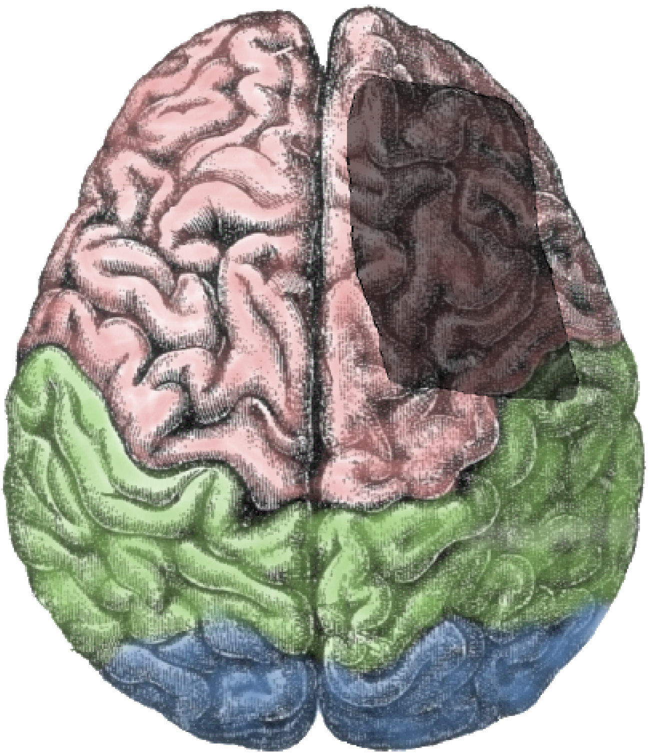
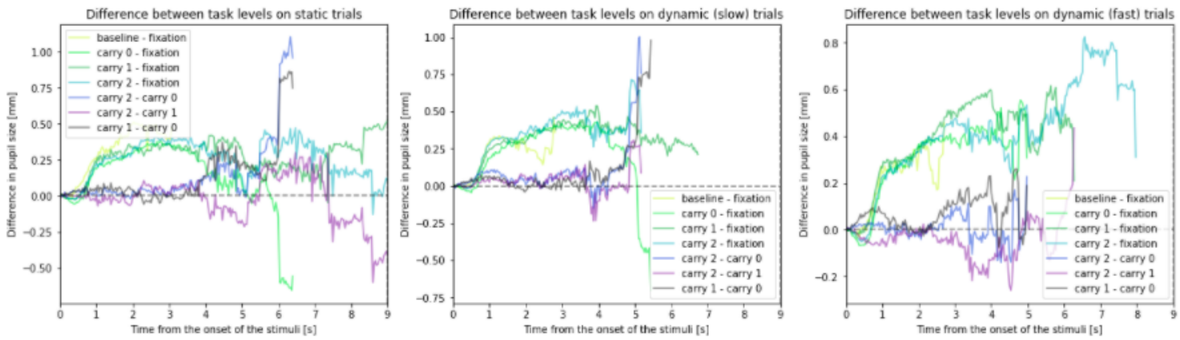
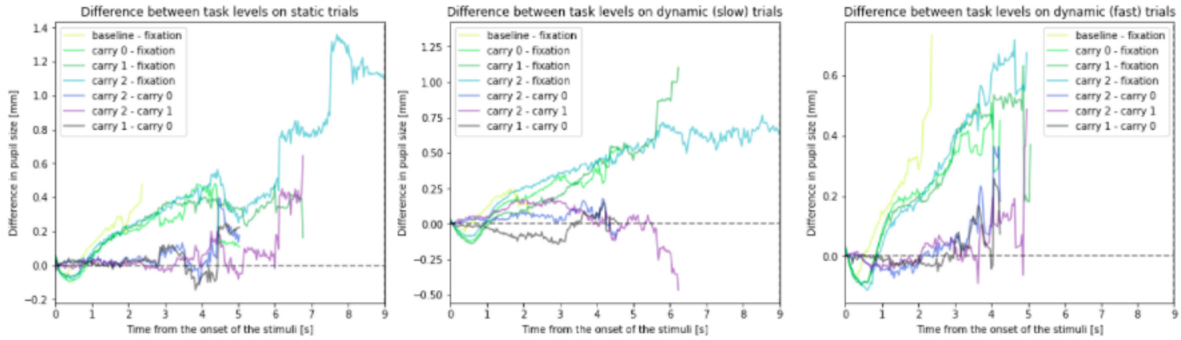


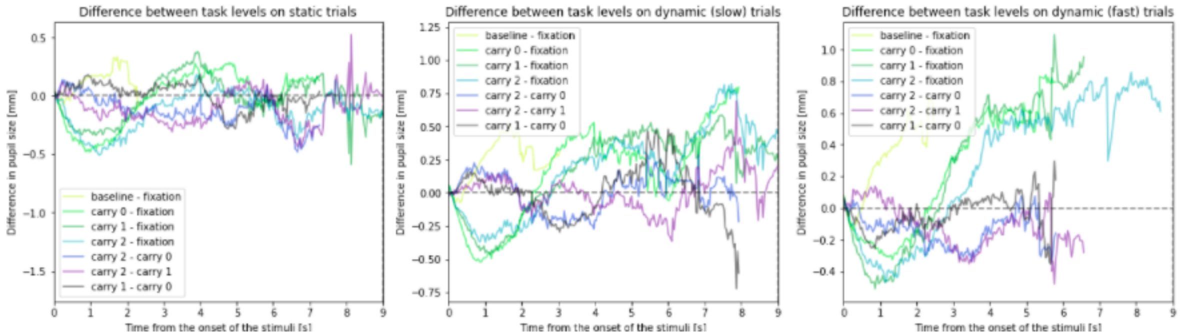
Figure 16: The approximate placement of the neuroimaging probe on the prefrontal cortex.



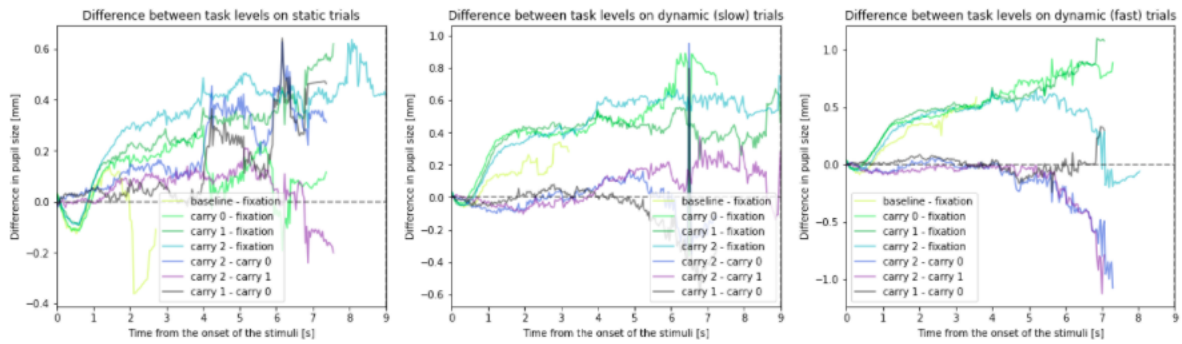
(a) Subject S2



(b) Subject J2



(c) Subject K2



(d) Subject R2

Figure 17: Trace differences in pupil size change for each of the subjects, plotted by presentation complexity.

5.3 Discussion

Cognitive load is the complex concept, sometimes not very well-specified. In the field of HCI it is defined as mental resources of the person available for solving problems or completing tasks at a time. Amount of cognitive load and attention required at any point can vary throughout the day. Often we are forced to divide attention and solve multiple tasks at the same time.

Adaptive user interfaces can provide appropriate support for these situations and can help avoid distractions and possibly dangerous situations. Such interfaces would provide adequate cognitive aid in the various environments.

Development of such systems requires real-time and precise information about cognitive load of the user. In this work the mental calculation tasks were used to find out whether cognitive load can be measured from physiological signals (hemodynamic brain activity and pupil size).

With careful design, valid assumptions and correct interpretation of the data it is possible to distinguish different levels of cognitive load from physiological measurements like hemodynamic brain activity and pupil size. When making inferences from the pupillary data, it is important to be aware of multiple noise sources, especially the luminance.

With the in-field measurements it is impossible to estimate luminance of the surroundings, and care should be taken on the analysis stage. There is a number of sophisticated methods for dealing with such issues, for example power spectrum analysis and general linear modeling methods, but the applicability of any method depends on the experimental design or the environment in which the data were collected.

Another important aspect in choosing the method for data analysis is the performance of the method and sensitivity to the data size. When designing an application, which would use the physiological measurements as the proxy for mental processes, one should think of the time demands in such application. The real-time performance is a challenging example which would require very powerful and reliable algorithms for dealing with the data. At the same time, in human-computer interaction and in designing the adaptive systems, the need for real-time applications is inevitable.

One contribution in the direction of the real-time analysis of physiological measurement was made by [37], who compared different psycho-physiological measures for accessing cognitive load and their real-time requirements; [38], who have user near-infrared spectroscopy for designing a passive brain-computer interface, and [3], who accessed cognitive load of young and older adults in real-time.

References

[1] Tan D., Nijholt A. (2010). Brain-Computer Interfaces and Human-Computer Interaction. *In: Tan D., Nijholt A. (eds) Brain-Computer Interfaces. Human-Computer Interaction Series. Springer, London*

[2] Riedl, R., Randolph, A., Brocke, J., Dimoka, A. (2010). The Potential of Neuroscience for Human-Computer Interaction Research. *The 9th Annual Pre-ICIS Workshop on HCI Research in MIS.*

[3] Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Roning, J., Forlizzi, J.F., Dey, A.K. (2014). Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. *IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014, pp. 39–48.* doi: 10.1002/hbm.21159.

[4] Fishburn, F. A., Norr, M. E., Medvedev, A. V., Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Frontiers in human neuroscience, 8: 76.* doi:10.3389/fnhum.2014.00076.

[5] Hossain, G., Yeasin, M. (2014). Understanding Effects of Cognitive Load from Pupillary Responses Using Hilbert Analytic Phase *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Conference.* doi:10.1109/CVPRW.2014.62.

[6] Klinger. J. (2010). Measuring cognitive load during visual tasks by combining pupillometry and eye tracking.

[7] Keshmiri, S., Sumioka, H., Yamazaki, R. and Ishiguro, H. (2017). A Non-parametric Approach to the Overall Estimate of Cognitive Load Using NIRS Time Series. *Frontiers in Human Neuroscience, 11:15.* doi:10.3389/fnhum.2017.00015.

[8] Verner, M., Herrmann, M. J., Troche, S. J., Roebers, C. M., Rammsayer, T. H. (2013). Cortical oxygen consumption in mental arithmetic as a function of task difficulty: a near-infrared spectroscopy approach. *Frontiers in Human Neuroscience, 7, 217..* doi:10.3389/fnhum.2013.00217.

[9] Purves, D., Augustine, G.J., Fitzpatrick, D. (2008). Neuroscience, Fourth Edition *Sinauer Associates, Inc., Hardcover..*

[10] Kotseruba, I., Tsotsos, J.K. (2016). A Review of 40 Years of Cognitive Architecture Research: Core Cognitive Abilities and Practical Applications.

[11] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12, 257-285.*

[12] Anderson, J. R. (1983). The Architecture of Cognition. *Cambridge, MA: Harvard University Press..*

[13] Ebbinghaus, H. (1885, translated 2003). Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences, 20(4), 155–156.*. doi:10.5214/ans.0972.7531.200408.

[14] Hebb, D. O., (1949) The Organization of Behavior: A Neuropsychological Theory *A Wiley book in clinical psychology Neuropsychological Theory.*

[15] Atkinson, R. C., Shiffrin, R. M. (1968). Chapter: Human memory: A proposed system and its control processes. *In Spence, K. W., Spence, J. T. The psychology of learning and motivation (Volume 2). New York: Academic Press. pp. 89–195.*.

[16] Baddeley, A. D., Hitch, G. (1974). Working memory. *In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47–89).*.

[17] Andres, M., Michaux, N., Pesenti, M. (2012). Common substrate for mental arithmetic and finger representation in the parietal cortex. *NeuroImage, 62 (3) pp. 1520-1528.* doi:10.1016/j.neuroimage.2012.05.047.

[18] Bear, M. F., Connors, B. W., Paradiso, M. A. (2016). Neuroscience: exploring the brain. *Philadelphia: Wolters Kluwer. Fourth edition.*

[19] Niederhauser, D. S., Reynolds, R. E., Salmen, D. J., Skolmoskii, P. (2000). Influence of cognitive load on learning from the hypertext. *Journal of Educational Computing Research 23(3).* doi:10.2190/81BG-RPDJ-9FA0-Q7PA.

[20] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63, 81-97.*.

[21] Mierlo, C. M., Jarodska, H., Kirschner, F., Kirschner, P. A. (2012). Cognitive Load Theory in E-Learning. *Encyclopedia of Cyber Behavior.* doi:10.4018/978-1-4666-0315-8.ch097.

[22] Pass, G. W. C., Van Merrienboer, J. J. (1994). Instructional control of cognitive load on the training of complex cognitive tasks. *Educational Psychology Review , 6(4).*

[23] Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity *Google Patents. US Patent 6090051.*

[24] Fischer, M. H. (2012). A hierarchical view of grounded, embodied, and situated numerical cognition. *Cogn. Process., 13 (1) pp. 161-164.* doi:10.1007/s10339-012-0477-5.

[25] Hirshfield, L. M. et. al. (2007). Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy. *ACM UIST 2007 Symposium on User Interface Software and Technology*

[26] Prado, j., Mutreja, R., Zhang, H., Mehta, R., Desroches, A. S., Minas, J.E. (2011). Distinct representations of subtraction and multiplication in the neural systems for numerosity and language. *Hum. Brain Mapp., 32 (11) pp. 1932-1947.* doi:10.1002/hbm.21159.

[27] Rosenberg-Lee, M., Chang, T. T., Young, C. B., Wu, S., Menon, V. (2011). Functional dissociations between four basic arithmetic operations in the human posterior parietal cortex: A cytoarchitectonic mapping study. *Neuropsychologia, 49(9), 2592–2608..* doi:10.1016/j.neuropsychologia.2011.04.035.

[28] Grabner, R. H., Reishofer, G., Koschutnig, K., Ebner, F. (2011). Brain Correlates of Mathematical Competence in Processing Mathematical Representations. *Frontiers in human neuroscience, 5:130.* doi:10.3389/fnhum.2011.00130.

[29] Tschentscher, N., Hauk, O. (2014). How are things adding up? Neural differences between arithmetic operations are due to general problem solving strategies *NeuroImage, 92:369-80.* doi:10.1016/j.neuroimage.2014.01.061.

[30] Imbo I., Vandierendonck, A., Rammelaere, S. (2007). The role of working memory in the carry operation of mental arithmetic: Number and value of the carry. *Q J Exp Psychol (Hove). 60(5):708-31..* doi:10.1080/17470210600762447.

[31] Hart, S. G., Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *In Hancock, P.A. and Meshkati, N. (Eds.), Human Mental Workload. Amsterdam, North-Holland, 139-183..*

[32] Zénon, A. (2017). Time-domain analysis for extracting fast-paced pupil responses. *Sci. Rep. 7, 41484.* doi: 10.1038/srep41484 doi:10.1038/srep41484.

[33] Hoeks, B., Levelt, W. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, Computers 25: 16..* doi:10.3758/BF03204445.

[34] Nissilä, I. et. al. (2017). Affective and non-affective touch evoke differential brain responses in 2-month-old infants. *NeuroImage.* doi: 10.1016/j.neuroimage.2017.12.024

[35] Ortiz, E. (2014). Optical Topography of Evoked Brain Activity during Mental Tasks Involving Whole Number Operations. *International Journal for Mathematics Teaching and Learning.*

[36] Rueckert, L., Lange, N., Partiot, A., Appollonio, I., Litvan, I., Le Bihan, D., Grafman, J. (1996). Visualizing Cortical Activation during Mental Calculation with Functional MRI. *Neuroimage. 3(2):97-103..* doi:10.1006/nimg.1996.0011

[37] Haapalainen, E., kim, S., Forlizzi, J. F., Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. *UbiComp '10 Proceedings of the 12th ACM international conference on Ubiquitous computing.* doi:10.1145/1864349.1864395.

[38] Girouard, A., Solovey, E. T., Jacob, R. J. K. (2013). Designing a Passive Brain Computer Interface using Real Time Classification of Functional Near-Infrared Spectroscopy. *International Journal of Autonomous and Adaptive Communications Systems, vol. 6, no. 1, pp. 26-44.* doi: 0.1504/IJAACS.2013.050689.